



UNIVERSIDAD DE LA REPÚBLICA
FACULTAD DE INGENIERÍA



Approaches to Interactive and Improvisational Storytelling

A THESIS SUBMITTED BY

Santiago Góngora De La Fuente

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF

MAGÍSTER EN INFORMÁTICA (PEDECIBA)

THESIS SUPERVISORS

Luis Chiruzzo Alonso Universidad de la República
Gonzalo Méndez Pozo Universidad Complutense de Madrid

JURY MEMBERS

Juan Diego Campo Universidad de la República, Uruguay
Lorena Etcheverry Universidad de la República, Uruguay
Lara J. Martin (Reviewer). University of Maryland, Baltimore County

ACADEMIC SUPERVISOR

Luis Chiruzzo Alonso Universidad de la República

Montevideo, Uruguay
August 11, 2025

Approaches to Interactive and Improvisational Storytelling, Santiago Góngora De La Fuente.

Pages: 166.

<https://www.pedeciba.edu.uy/es/area/informatica/>

Sprouts of time
The roots evolving
The seeds for the future were scattered yesterday
Tomorrow's harvest field we plant today

ANGRA

This page was intentionally left blank.

Acknowledgements

The path that led to this document was very long (it began many years before I started my Master's thesis!), and many people supported me along the way, so there are many to acknowledge in this section. I will begin by thanking the Uruguayan public education system, which has given me many opportunities throughout the years, and also gave my parents the opportunity to educate their son. In this regard, thanks to PEDECIBA for providing a high-quality educational program that helps shape the future of our science and the whole educational system. I also wish to thank ANII for granting me a scholarship (POS_NAC_2022.1.173659) that enabled me to remain focused on this thesis.

Special thanks to all my friends, many of whom share my passion for storytelling, video games and boardgames. First, thank to those with whom I have played TTRPGs: Jojo, Masi, Maxi, Renzi, Goyco, Bru, Gio and Tiago. Thank you as well to all my friends with whom I have spent many hours playing board and video games: NicoG, AgusP, Cane, AgusC, Guille, Facu, Jenny, NicoE, Dami and Sofi.

Thank you to all the special people I have worked with who have become very important to me. I'll start by thanking all the members of my lab, Grupo PLN: those who have mentored me over the years (Dina, Aiala, Diego, Guille and JJ), those fellow undergraduate and graduate students with whom I shared many seminars and discussions (Santi, Mathi, Nachito, Guille and NachoR), and also those who have joined us in recent months. Many of these colleagues, along with several of my personal friends I mentioned earlier, took part in the evaluation of the PAYADOR system, so thank you once again. Also, thanks to other colleagues+friends that are not part of our lab, but have a special role in this process: Camily, Belu, Leíto and Trini, you are definitively not hardcore gamers, but you showed me there are many other interesting topics to talk about! (jiji). Thanks to all my colleagues who chatted with me about this over lunch and in seminars, and a special thanks to Anita for the wonderful story I included in Chapter 2. Finally, thanks to Adri, Flavia, FDO and SebaG: you were also very important catalysts for my academic achievements!.

I would like to give special thanks to some truly important people. First, thank you, Luis Chiruzzo, for sharing your knowledge with me and guiding me throughout this process, as well as for sharing deep important interests, like how and when growls, tenors and sopranos are perfectly blended in a musical piece. Second, thanks to Gonzalo Méndez and Pablo Gervás for supporting Luis in this task, making significant contributions, being very generous with this (initially) random Uruguayan guy, and warmly welcoming me to *la Complutense* in 2023. Third, thank you Lara J. Martin, for agreeing to review this thesis and suggesting improvements that made it a much better document; I cannot put into words how much this means to me. Thanks also

to Lorena Etcheverry and Juan Diego Campo for being the other two members of the jury, asking me really interesting questions from your unique perspectives, which will surely fuel my future work. Fourth, I am grateful to several researchers who have generously shared their thoughts on my work: Luciana Benotti, Laura Alonso Alemany, Jocelyn Dunstan, Maite Martín, Álvaro Cabana, Thamar Solorio, Rafael Pérez y Pérez and Dan Brown. Lastly, and with special gratitude, I want to thank again three people I have admired deeply even before beginning work on this thesis: Pablo Gervás, Lara J. Martin and Luciana Benotti. Thanks for the inspiration; having this thesis enriched by your comments is a truly meaningful achievement in my life.

Finally, and most especially, I would like to give special thanks to Agustina, my wife, with whom I spent many hours enjoying and studying games, stories, and everything in between. We started dating as undergraduates back in 2016, talking about wizards and all sorts of geeky stuff, and, in the middle of my Master's years, we got married, filling our house with many — many— board games. And, of course, thanks to my parents, Alicia and Javier — and their families — who laid the foundations for my life's projects, including my academic ones.

Gracias a todas y todos.

PS: Thanks to all the artists I listened to while working on this thesis.

*To my mother, and her family, for showing me the
importance of the arts to understand the world.*

*To my father, and his family, for showing me the
importance of storytelling to be in touch with reality.*

This page was intentionally left blank.

Abstract

In Tabletop Role-playing Games, players act as characters living in an ever-evolving fictional world. Usually, among the players there is one with a special role, the Game Master (GM), who is in charge of bringing the fictional world to life, acting as every living being the players meet, narrating every scene, and describing how the world changes after the players' actions. In order to achieve this, the GM relies on narrative improvisation to generate a fertile space of collaborative narrative creation. During the last decades, several computational approaches have been proposed to model this narrative experience, being "Interactive Fiction" games (e.g. the *Zork* series) and those called "Computer Role-playing Games" (e.g. the *Baldur's Gate* series), the most famous among them. Since they work by pre-programming a great amount of pre-authored content, the modeling of *improvisational storytellers* (e.g. GMs) has been little explored until the advent of Large Language Models (LLMs), at the beginning of the 2020s decade.

This document presents some approaches to build and evaluate *Interactive Narrative* systems that exhibit some improvisational abilities. First, we propose three tests to evaluate gamemastering models and use them to evaluate three LLMs: ChatGPT, Bard, and OpenAssistant. By focusing on specific actions, these tests try to limit the usual variability present when evaluating open-world narrative systems. In the experiments we find that ChatGPT and Bard can deliver a satisfying role-playing experience, but exhibiting flaws in updating the world state after the player actions. In contrast, OpenAssistant struggles maintaining the GM role during the session.

Secondly, we present **PAYADOR**, a gamemastering model consisting in an LLM grounded on a symbolical representation of the fictional world. This way, the LLM has two main responsibilities: (1) to narratively describe the world based on that representation, and (2) to suggest changes to it, after processing the player actions. In order to do the latter, the approach is based on modeling *transformations* the fictional world can have, what means a change of focus away from the classical strategy of modeling individual player actions. To evaluate PAYADOR, both in Spanish and English, eight people played two scenarios with different evaluation objectives. We also wanted to evaluate the approach with different LLMs, so half of them used Llama and half of them used Gemini. All the evaluators were able to finish both scenarios, what seems to indicate the approach is viable and fully functional. After analyzing their playthroughs, we find that *transformations* are general enough to allow players use their creativity to overcome the challenges present in the scenario. Even when LLMs may fail, the symbolic system allowed us to understand what the errors were and reflect on their causes. Based on those errors, we discuss how to improve PAYADOR and also reflect on the following steps towards modeling the phenomenon of Improvisational Storytelling in Open Worlds. The playthroughs of the eight evaluators for both scenarios, and the version of PAYADOR used in those experiments — that considers the *Moved Items*, *Unblocked locations* and *Player movement* transformations — are available on GitHub.

Resumen

En los Juegos de rol de mesa, los jugadores actúan como personajes que viven en un mundo ficticio en constante cambio. Normalmente, entre esos jugadores hay uno con un rol especial, el *Director de juego* (GM, por sus siglas en inglés), quien está a cargo de darle vida al mundo ficticio, actuando como cada ser vivo que los jugadores se encuentran, narrando cada escena, y describiendo cómo cambia el mundo a raíz de las acciones de los jugadores. Para lograr esto, el GM recurre a la improvisación narrativa para generar un espacio fértil de creación narrativa colaborativa. En las últimas décadas se han propuesto varios enfoques para modelar computacionalmente esta experiencia narrativa, siendo “Interactive Fiction” (por ej., la serie *Zork*) y “Computer Role-playing Games” (por ej., la serie *Baldur’s Gate*) los tipos de juegos más famosos entre ellos. Como esos enfoques funcionan mediante programar de antemano un montón de contenido escrito previamente, el modelado de *narradores improvisacionales* (por ej., GMs) ha sido poco explorado hasta la aparición de los Grandes Modelos de Lenguaje (LLM, por sus siglas en inglés), al inicio de la década de los 2020s.

Este documento presenta varios enfoques para construir y evaluar sistemas de *Narrativa Interactiva* que muestren algunas habilidades de improvisación narrativa. Primero, proponemos tres *tests* para evaluar modelos de GMs y los usamos para evaluar tres LLMs: ChatGPT, Bard, y OpenAssistant. Enfocándose en acciones específicas, estos *tests* intentan limitar la variabilidad habitual que está presente al evaluar sistemas narrativos de mundo abierto. En los experimentos encontramos que tanto ChatGPT como Bard pueden ofrecer una experiencia agradable, pero mostrando defectos al actualizar el estado del mundo a partir de las acciones del jugador. En cambio, OpenAssistant muestra problemas en mantener el rol de GM durante la sesión.

Segundo, presentamos **PAYADOR**, un modelo de GM que consiste en hacer *grounding* a un LLM en una representación simbólica del mundo ficticio. De esta manera, el LLM tiene dos responsabilidades principales: (1) describir narrativamente el mundo basándose en esa representación, y (2) sugerir cambios a hacerle, a partir de las acciones del jugador. Para hacer lo segundo, este enfoque se centra en modelar *transformaciones* que el mundo puede tener, lo que significa un cambio de foco con respecto a la estrategia clásica de modelar acciones individuales que el jugador puede hacer. Para evaluar PAYADOR, tanto en inglés como en castellano, ocho personas jugaron dos escenarios con diferentes objetivos de evaluación. También queríamos evaluar el enfoque con diferentes LLMs, así que una mitad de las personas usaron Llama y la otra mitad usó Gemini. Todos los evaluadores pudieron terminar los escenarios, lo que parece indicar que el enfoque es viable y totalmente funcional. Luego de analizar sus trazas de juego, encontramos que las *transformaciones* son lo suficientemente generales para permitir que los jugadores usen su creatividad para superar los desafíos del escenario. Incluso aunque los LLMs pueden fallar, el sistema simbólico nos permitió entender cuáles fueron los errores y reflexionar sobre sus causas. Basándonos en esos errores, discutimos cómo mejorar PAYADOR y también reflexionamos sobre cuáles son los siguientes pasos hacia el modelado de la improvisación narrativa en mundos abiertos. Las trazas de juego de las ocho personas que jugaron los dos escenarios, y la versión de PAYADOR usada en esos experimentos — que considera las transformaciones *Moved Items* (objetos movidos), *Unblocked locations* (lugares desbloqueados) y *Player movement* (movimiento del jugador) — están disponibles en GitHub.

Contents

| | |
|---|------------|
| Acknowledgements | iii |
| 1 Introduction | 1 |
| 1.1 Objectives | 3 |
| 1.2 Contributions | 3 |
| 1.3 Structure of this document | 4 |
| 2 The State and Fate of Interactive and Improvisational Storytelling | 5 |
| 2.1 Interactive and Improvisational Storytellers | 5 |
| 2.1.1 Interactivity — A book full of branches | 6 |
| 2.1.2 Improvisation — A bedtime story | 7 |
| 2.1.3 Tabletop Role-playing Games: Interactive and improvisational storytelling with rules | 8 |
| 2.2 Computer Role-playing Games: The first gamemastering models | 10 |
| 2.2.1 A natural language interface — Zork | 10 |
| 2.2.2 A huge closed world — Baldur’s Gate 3 | 12 |
| 2.3 Natural Language Processing | 15 |
| 2.3.1 Rule-based models | 15 |
| 2.3.2 Large Language Models | 16 |
| 2.4 NLP-powered models of Role-playing Games | 18 |
| 2.4.1 Assistants | 19 |
| 2.4.2 Datasets and Environments | 19 |
| 2.4.3 Generation of content | 20 |
| 2.4.4 Other gamemastering mechanisms | 21 |
| 2.4.5 General gamemastering models | 21 |
| 2.4.6 AI Dungeon | 22 |
| 2.5 Closing remarks and key points | 23 |
| 3 In Quest for the Neuro-symbolic Truth | 25 |
| 3.1 Taking subjectivity out of the equation | 26 |
| 3.1.1 Item tracking | 27 |
| 3.1.2 Map consistency | 28 |
| 3.1.3 GM-P-GM | 28 |
| 3.2 Experiments with three LLMs | 29 |
| 3.2.1 Quantitative results | 30 |
| 3.2.2 Four observations on this interactive experience | 31 |
| 3.3 The Neuro-symbolic truth | 33 |
| 3.4 Closing remarks and key points | 35 |

Contents

| | | |
|----------|--|------------|
| 4 | PAYADOR: A Neuro-symbolic Gamemastering Model | 37 |
| 4.1 | The world and its components | 38 |
| 4.2 | A change of focus for the world-update process | 39 |
| 4.2.1 | But... what is the difference? | 41 |
| 4.3 | Prompting Large Language Models | 41 |
| 4.3.1 | World rendering | 42 |
| 4.3.2 | LLMs to describe the scene | 43 |
| 4.3.3 | LLMs to predict the effects of the user input | 45 |
| 4.4 | Consistent updates in the world | 46 |
| 4.5 | Overview of the PAYADOR approach | 47 |
| 4.5.1 | 5 steps in a loop | 48 |
| 4.5.2 | Some comments on the game loop | 50 |
| 4.6 | Closing remarks and key points | 52 |
| 5 | An Empirical Evaluation of PAYADOR | 55 |
| 5.1 | The evaluation plan | 56 |
| 5.1.1 | Adjust the focus | 56 |
| 5.1.2 | The app | 57 |
| 5.1.3 | The people | 58 |
| 5.2 | Scenario A: Emma and her turtle | 58 |
| 5.2.1 | Milestones and evaluation goals | 59 |
| 5.2.2 | Quantitative results | 61 |
| 5.2.3 | Qualitative observations | 64 |
| 5.3 | Scenario B: Venancio looking for Artigas | 69 |
| 5.3.1 | Milestones and evaluation goals | 70 |
| 5.3.2 | Quantitative results | 71 |
| 5.3.3 | Qualitative observations | 73 |
| 5.4 | Overall findings and future work | 78 |
| 5.4.1 | Transformations | 79 |
| 5.4.2 | Fog of war | 80 |
| 5.4.3 | Lack of memory | 81 |
| 5.4.4 | Narrated outcomes and scene descriptions | 82 |
| 5.4.5 | PAYADOR and LLMs in the co-creation process | 84 |
| 5.5 | Closing remarks and key points | 84 |
| 6 | Conclusions | 87 |
| 6.1 | Contributions of this thesis | 88 |
| 6.2 | Future work | 89 |
| 6.2.1 | Observations extracted from the evaluation | 89 |
| 6.2.2 | Additional features to complement PAYADOR | 90 |
| 6.2.3 | The tip of the gamemastering iceberg | 91 |
| 6.3 | Closing remarks | 92 |
| | References | 93 |
| | Appendices | 109 |
| A | Prompts to evaluate LLMs as out-of-the-box game masters | 111 |
| A.1 | GM-P-GM | 111 |
| A.2 | Item tracking | 112 |
| A.3 | Map consistency | 113 |

| | |
|--|------------|
| B PAYADOR: Prompts for LLMs | 115 |
| B.1 Narrate the scene | 115 |
| B.1.1 English | 115 |
| B.1.2 Spanish | 116 |
| B.2 Predict outcomes of the user input | 117 |
| B.2.1 English | 117 |
| B.2.2 Spanish | 119 |
| C Playthroughs of PAYADOR | 123 |
| C.1 Scenario A | 123 |
| C.1.1 Gemini — TesterB | 123 |
| C.1.2 LLama — TesterC | 131 |
| C.2 Scenario B | 138 |
| C.2.1 Gemini — TesterA | 138 |
| C.2.2 LLama — TesterD | 141 |
| D Commented playthroughs of PAYADOR | 147 |
| D.1 Scenario A | 147 |
| D.1.1 TesterG struggling with the <i>lack of memory</i> of PAYADOR . . | 147 |
| D.1.2 TesterG in a co-creative act | 148 |
| D.2 Scenario B | 149 |
| D.2.1 TesterH and the <i>communication chaos</i> | 149 |

This page was intentionally left blank.

Chapter 1

Introduction

When I was 15 years old, a classmate told me about “Role-playing games”. “Characters”, “levels”, “spells”, “combat rounds”... during the first minutes of the conversation I was convinced he was talking about a video game. Through *collaborative grounding* (Benotti and Blackburn, 2021) (i.e. discussing about it until we reached a shared agreement) I immediately realized that he was actually talking about a tabletop game. My confusion was reasonable, since the influence of Tabletop Role-playing Games (TTRPG) in video games was huge (MacCallum-Stewart et al., 2018), and many *video game concepts* my classmate was mentioning are examples of that remarkable influence.

TTRPGs are a type of game where players sit around a table (or, nowadays, join a voice chat in Discord) and tell a story together (Hitchens and Drachen, 2008). Although they are very diverse, and several variants have been proposed throughout the years, most TTRPGs consist of players acting as fictional characters in a fictional world. Among them, there is usually a special player, the **Game Master** (GM), who is in charge of describing the fictional world for the players, narrating the changes in the world as a consequence of their actions, creating challenges to keep them engaged in the game, and also acting as a judge, applying the rules and moderating related discussions. In Figure 1.1 there is a picture of a TTRPG group during a session, where four players are looking at the Game Master, who is probably explaining some rules, the outcome of the players’ actions or describing a new scene.

What TTRPGs did back in the 1970s was to formalize something that humans have been doing for several thousand years. We — humans — have always told stories to communicate with other humans, and that has been part of our evolution as a species (Boyd, 2018). We know how to tell them. We also know what it is to *improvise* a story, or some parts of it. And definitively we know what is to tell a story in company of another storyteller. However, what we did not do before TTRPGs was to set rules to play games based on those communicative acts. Some of these rules may seek to make the gameplay balanced, just like rules work in classic board games, but the difference lies in all the other rules, such as those dedicated to specify how characters can make the fictional world change, how characters can convince non-playable characters (i.e. those controlled by the GM) of something, and how all the living beings that inhabitate the fictional world can change (e.g. learning new abilities, changing their mood, and so on).

Chapter 1. Introduction



Figure 1.1: Five people playing a TTRPG session. The Game Master is the person sitting at the end of the table, protecting the game secrets with a *game master screen*. The image was taken from *Wikimedia commons*.

Throughout the years many efforts have been done to adapt the original TTRPG experience to video games and consequently many video game genres were born. First, *Text games* focused on the GameMaster-Player interaction, narrating scenes and describing the outcomes of the player actions (Reed, 2023). Then, *Computer Role-playing Games* (CRPG) focused on modeling complex character-development systems and challenging the player to think strategies to manage resources like food, magic, health and so on. Then, *Massively multiplayer Online Role-playing Games* tried to take CRPGs to the next level, having a lot of human players acting in a fictional world, at the same time, in any part of the world (Adams, 2009). However, all these adaptations could not deal with the most important aspect in TTRPGs: the **free will** of the players.

Modern games, like Baldur's Gate 3, have tried to tackle this problem by preprogramming many — many — actions that players can take, what really gives a greater illusion of free will. But that is not enough. TTRPGs are about creativity and improvisation, about acting as another living being, about taking coherent decisions with that identity. It is really unlikely that a finite set of preprogrammed actions will be enough to represent all what we can think of. We — humans — first needed language to fully unlock our creativity and narrative skills (Boyd, 2018), so it is natural to think that the path towards modeling the TTRPG experience will have to consider techniques to process the languages we speak to create stories.

All these type of systems, that let the user decide how the story will unfold, are called *Interactive Storytelling* (or *Interactive Narrative*) systems. Depending on the degree of agency the user has, the delivered experience can be similar to reading book with few different endings, similar to playing a TTRPG, and everything in between. The degree of agency that players have when playing a TTRPG was previously concep-

tualized by Martin et al. (2016) as *Open-world Improvisational Storytelling*. Naturally, since the best way to express our creative ideas is using our language, this field is intimately related to the field of Natural Language Processing (NLP), mainly concerned in building tools that can process and generate natural languages.

1.1 Objectives

The main objective of this thesis is to explore what approaches are suitable to build Open-world **Improvisational** and **Interactive** Storytelling systems, with a focus on the opportunities that Large Language Models (LLMs) have brought so far. More specifically, we want to explore how LLMs could be used to model some of the abilities that a human GM needs to run a TTRPG session. We mainly focus on two gamemastering abilities: tracking the world state consistently (i.e. without introducing details that contradict previous world states or contradict rules that govern the fictional world), narrating events, describing scenes, and answering clarification questions about the fictional world. Other gamemastering abilities, like introducing stochasticity from dice rolls, improvising new content on the fly, or engaging in deep dialogues, are not in the scope of this thesis.

Going into detail, the objectives of this thesis are as follows:

- To explore what methodologies can be followed to evaluate gamemastering models.
- To check if LLMs can work as out-of-the-box gamemastering models, carefully analyzing the playthroughs to determine what are their virtues and flaws.
- To discuss the potential of neuro-symbolic approaches to make gamemastering models more stable in terms of narrative coherence.
- To propose a neuro-symbolic approach to model some of the abilities that a human GM needs to run a TTRPG session, trying to enhance the capabilities of LLMs while also having the possibility to detect some of common errors they make when dealing with Improvisational Storytelling.
- To use that approach to build an interactive system that work as a — small but functional — gamemastering model.
- To evaluate the aforementioned approach through having humans playing with the interactive system, carefully analyzing the playthroughs to determine what are their virtues and flaws.

1.2 Contributions

The work presented in this thesis was done in the span of two years, from December 2022 to December 2024. During that time, we published some papers describing some parts of this work. The specific contributions of this thesis are as follows:

- The paper “Skill check: Some considerations on the evaluation of gamemastering models for role-playing games” describes a novel methodology to evaluate LLMs working as out-of-the-box gamemastering models. I am the first author of the paper, which I presented in Dublin (Republic of Ireland) at the 12th International Conference on Games and Learning Alliance, and was published in Springer Lecture Notes in Computer Science (Góngora et al., 2023).

Chapter 1. Introduction

- The paper “PAYADOR: A Minimalist Approach to Grounding Language Models on Structured Data for Interactive Storytelling and Role-playing Games” describes a novel approach to the world-update problem in Interactive Narrative. I am the first author of the paper, which I presented in Jönköping (Sweden) at the 15th International Conference on Computational Creativity, and was published in the proceedings of the conference (Góngora et al., 2024).
- The **PAYADOR** system, described in the aforementioned paper, is available on GitHub as functional code¹. It can be used as a starting point for future research in Interactive Storytelling.
- The last version of **PAYADOR** — the one presented in this thesis and used to perform the experiments — is available on GitHub².

I also wrote a small piece to submit to the *Early Career Symposium* that took place in Jönköping (Sweden) before the 15th International Conference on Computational Creativity. It was accepted and it is available on the web of the conference with the name of “Exploring approaches to open-world Interactive Storytelling”³ (Góngora, 2024).

1.3 Structure of this document

The rest of this document is structured as follows.

Chapter 2 defines all the concepts related to the problem of modeling a Game Master and those needed to understand the work presented in this thesis. It also presents previous work related to modeling some aspects of TTRPGs.

Chapter 3 presents an evaluation strategy to test LLMs as out-of-the-box gamemastering models, and a discussion on some considerations that should be taken into account when building gamemastering models.

Chapter 4 presents **PAYADOR**, a neuro-symbolic gamemastering model we propose, introducing the game loop it implements and the *transformations* it relies on in order to tackle the *world-update* problem in Interactive Storytelling.

Chapter 5 presents the methodology we followed to evaluate **PAYADOR**, the results we obtained, and a discussion on what to do next to improve the approach.

Finally, Chapter 6 presents the conclusions of the work presented in this document, summarizing our work, and also discussing some problems that may be further explored in the future.

Throughout this document I will use two pronouns, both including me. By default, I will use “I” to communicate personal anecdotes, opinions, and refer to the writing of the document. I will use “we” to describe the work I did with my advisors, Luis Chiruzzo and Gonzalo Méndez Pozo. Pablo Gervás Gómez-Navarro helped us to conceptualize many of the work presented here, so in many cases he will also be part of the group of people indicated by the “we” pronoun.

¹<https://github.com/pln-fing-udelar/payador>

²<https://github.com/sgongora27/msc-thesis-2025>

³https://computationalcreativity.net/iccc24/wp-content/uploads/2023/12/Gongora_ECS_ICCC24.pdf

Chapter 2

The State and Fate of Interactive and Improvisational Storytelling

The field of Interactive Narrative is an interdisciplinary one, where research can be focused on different aspects of human-computer interaction when it comes to telling stories. Therefore, in order to clarify the scope and terminology of the work presented in this thesis, in this chapter I will cover the main concepts that are relevant to contextualize both its conception and development.

2.1 Interactive and Improvisational Storytellers

When we think about *storytellers* some images might cross our minds. Depending on our personal background, we can probably imagine a parent reading a book to their children before bed, or even a bunch of people carefully listening to someone telling a story about something that may (or not) have happened some time ago, as represented by Farbman's photography shown in Figure 2.1. If someone asked me what my father's family meetings were like when I was a child, I would probably have answered that everyone was near a table eating *tortas fritas*, drinking mate and telling stories!. As an adult, I became aware that this specific tradition was just a mechanism to catch up with the loved ones. And now, as I write this thesis, I know that was totally expected, since humans are innate storytellers and they use stories to communicate diverse ways of meaning (Schiff, 2012).

In most of those cases, the story being told is linear, what implies that which event goes after which other event was decided — as a matter of historical veracity or just creativity — before the storyteller writes or tells them (Trichopoulos et al., 2023). This way, the listeners or readers



Figure 2.1: As taken from LIFE magazine: "Bushman children sitting around their chief as he acts out a story, southern Kalahari Desert in central-southern Africa. (Photo by N.R. Farbman/The LIFE Picture Collection © Meredith Corporation)".

Chapter 2. The State and Fate of Interactive and Improvisational Storytelling

have a passive role, in which they unveil the mysteries behind the plot without the power of changing anything in it. I tried to be careful about the previous words, since *linear storytelling* does not mean that the *storyteller* is the only one with the power of creating that story, nor it means that the story was fully planned beforehand. In the former case, the story might be **collaboratively** created by a group of writers (e.g. a movie plot), in which the *storyteller* could be part of it, although it is not mandatory (e.g. while reading a book aloud). In the latter case, the *storyteller* might have planned some details (or none) of the story, while the rest of them will be **improvised** during narration, such as happens in improvisational theater (Martin et al., 2016). These are core concepts for the work presented here, and I will come back to them soon in this chapter.

Furthermore, we — humans — created an even more marvelous thing: **interactive** storytelling. From then on, people can actually *interact* with the storyteller(s) and have *some* degree of control over the story being unfolded (Aylett and Louchart, 2007; Riedl and Bulitko, 2012). *How much* degree of control will strictly depend on the capabilities of the media used to tell the story and the decision of the storyteller: the more degree of control one has, the less the other has, what is known as the Interactive Dilemma (Peinado and Gervás, 2004) or Narrative Paradox (Aylett, 1999; Aylett and Louchart, 2007).

Since interactivity is a very powerful tool for enhancing storytelling (Cavazza and Pizzi, 2006; Hargood et al., 2022), it has been used long before computers became widely available to the general public. From books to board games, interactivity played a key role in the development of present-day entertainment. In the following sections I will describe some illustrative examples for Interactive or Improvisational Storytelling, trying to clearly define the concepts that are key to understand the problem of modeling some of the skills that a Game Master needs in order to run an RPG session, which is aligned to what Martin et al. (2016) defined as *Open-world Improvisational Storytelling*.

2.1.1 Interactivity — A book full of branches

Perhaps the most clear example of non-digital Interactive storytelling are the “Choose Your Own Adventure” (CYOA) book series. In these books, the writer(s) narrate to the reader a story that was previously written in a modular way: after reading a page, the reader has one or more options, resulting in different pages to read.

Since this is systematic, from the beginning to the end of the story, the narrative structure of these books can be represented as a directed graph where each node is linked to one or more nodes, as shown in Figure 2.3. Each of these nodes can be thought as a reusable unit and, in fact, this is the main idea behind present-day interactive storytelling: the designer creates a story, previously divided in modules, that can be concatenated in order to get many different stories. This powerful idea is represented by the classic orange circle included in the front cover of the series (see Figure 2.2), stating how many different endings the reader can get.

The type of experience delivered by systems that work this way is called *branching narrative* (Aylett and Louchart, 2007). In this type of system, the *user agency* (the free will of the player) is reduced to a minimum (Riedl and Bulitko, 2012). It is not only that the options the user has in each node are just a few, but also that the granularity of the fictional world (Arjoranta, 2017) is rigidly set beforehand,

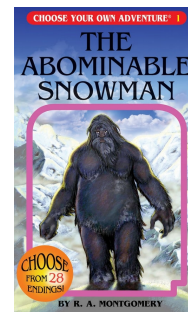


Figure 2.2: The front cover of “The Abominable Snowman” book (Montgomery, 2007).

2.1. Interactive and Improvisational Storytellers

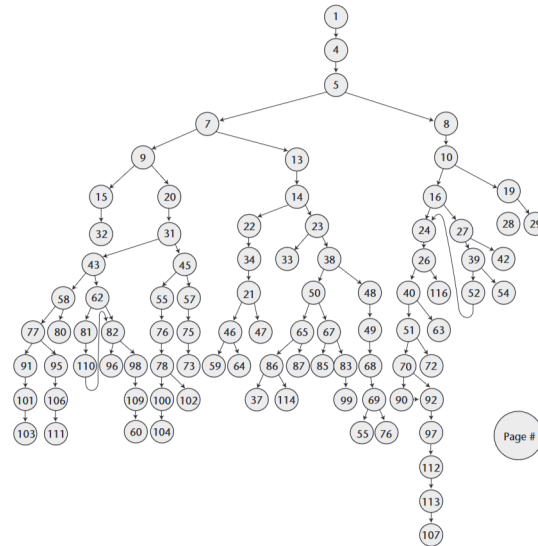


Figure 2.3: A Graph representation for “The Abominable Snowman” book (Montgomery, 2007), directly taken from Riedl and Bulitko (2012).

what prevents the reader to dig further in some things that may seem interesting (e.g. knowing the color of an object mentioned while describing the scene). As I will explain in section 2.2, this problem was tackled by taking advantage of the great storage capacity of modern computers, where the amount of nodes in the graph can grow to a really high number, making the users feel their agency is unlimited.

2.1.2 Improvisation — A bedtime story

The CYOA type of systems offer the user some degree of interactivity. What they do not do — at all — is to improvise: the pages are already written, and there is no chance to include different content.

However, improvisation is an intrinsic characteristic of humankind, consisting of inner rules to cope with problems for which we do not have previously internalized rules (Bertinetto and Bertram, 2020). We improvise when we run out of some ingredient for our food, we improvise when we do not know what to say to someone, we improvise when we want to do something in a different way, and we improvise when something does not go as expected. It is also a needed tool both for art creation and live performances (Bresnahan, 2015) and, as expected, storytelling is not an exception: from an early age, improvisation plays a key role while creating stories (Sawyer, 2002). Therefore, improvisation is a necessary condition to achieve a high degree of role-playing.

While writing this thesis, one day I was having lunch with my workmates and one of them shared with us a bedtime story that was a classic in her family. It tells the story of a kid wanting to play with “Mr. Dream”, looking for him in different places or objects, and not being able to find him. After some time, the storyteller begins to tell the story again, continuing the loop, varying the places and objects each time, until the kid falls asleep. I asked my workmate to include the story in this thesis, and a transcription is shown in Figure 2.4, both in Spanish (the original one she was told) and English.

Chapter 2. The State and Fate of Interactive and Improvisational Storytelling

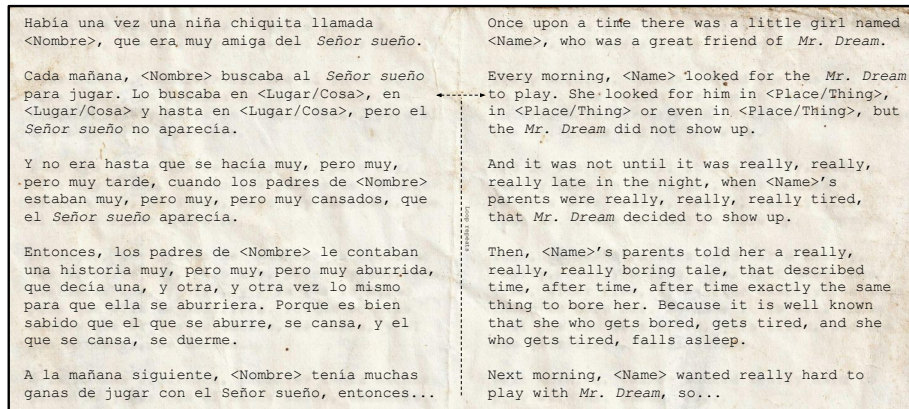


Figure 2.4: A bedtime story that can be told with different variations each time. The loop is indicated by a dashed arrow. On the left, the story in Spanish; on the right, in English.

The beauty of this story is that it has a rigid structure but, at the same time, it lets the storyteller to choose places or objects that sound familiar to the kid. It is important to notice that the improvised content can be anything, which implies that the story, as a whole, will likely not lose coherence whatever places or objects the storyteller decides to include. Thinking of it as a storytelling system, where the places or objects are chosen randomly or following an if-then strategy, it would exhibit some ability to improvise. Finally, this story is an illustrative (and not dependent on digital media) example of how present-day RPG video games work: the designers write template dialogues that are then instantiated with the name of the characters, or other relevant data saved in the game's database.

Now that I have introduced the concepts of **interactivity** and **improvisation**, let's dive into the main topic of this thesis: Role-Playing Games.

2.1.3 Tabletop Role-playing Games: Interactive and improvisational storytelling with rules

In 2016, the Netflix's *Stranger Things* TV show came out, proposing to revisit some of the nostalgic corners of the US culture (Hassler-Forest, 2020), and, by extension, of the world's culture. Among all the things that the series pays homage to, Role-playing Games were one of the main components of the formula. Not only the script was full of references to the hobby itself (see Figure 2.5), but also it was written in a style (Bastow, 2021) typical of *Dungeons and Dragons* (D&D), the first Role-playing Game to be published (Mearls et al., 2018; Ewalt, 2013).

This focus in D&D is totally reasonable, since back in the 1970s and 1980s, the boardgaming community was amazed by the first versions of the game, born as a fork of war games (Ewalt, 2013). Instead of managing an army, each player plays the role of a single character acting in a **fictional world**, having the possibility to take decisions and interact with it. Also, the characters have skills that make them useful for the needs of the adventuring group, also known as *the party* (Mearls et al., 2018). These revolutionary ideas laid the foundations of Role-playing games, but it was not enough, and soon after they turned into an innovation laboratory: statistics to measure the character's skills, that could increase according to experience; rolling dice to determine the outcomes of a specific effort according to the character's skills;

2.1. Interactive and Improvisational Storytellers



Figure 2.5: Some of the main characters from Stranger Things playing Dungeons & Dragons. This frame was directly taken from the first episode of the series on Netflix.

players having to act in-character, as in a theater play, bringing them to life based on a rich background story (Ewalt, 2013); and other groundbreaking game mechanics. The complexity of the games increased, and they were later considered as a form of organized collaborative storytelling (Katifori et al., 2022) through dialogue (Mäyrä, 2017), similar as what happens in Improvisational Theater (Durall and Perrin, 2023).

Among all those novel concepts, the one that we focus on in this thesis is the Game Master (GM), a special type of player. The Game Master (Tychsen et al., 2005) stands out as the one typically in charge of orchestrating the game, narrating the scenes and describing the places the players will explore, bringing to life the non-playable characters (NPCs) they will meet there and also designing the challenges that, collectively, all of them will face together. Although GMs usually plan the adventures in advance, sometimes using an if-then strategy (Acharya et al., 2021c), one of the main differences between RPGs and the branching narrative systems described in Section 2.1.1 is that the story being co-created by all the players takes place in an **open world**, a fictional place they are free to explore with little to no restrictions (Tychsen et al., 2005). This implies that the GM must frequently improvise to cope with coherence problems that arise due to the free will of the players (Strugnell et al., 2018), choosing the best way to narrate each outcome (Flowers et al., 2006), and deciding which rules and how to apply them (Durall and Perrin, 2023), in order to prevent the players to experience a feeling of unfairness (Katifori et al., 2024). For that reason, the GM is also considered a judge that tries to ensure the consistency of the game, as a way to provide a good experience to the players (Durall and Perrin, 2023).

The previously described gamemastering responsibilities are classic in games like D&D and *Call of Cthulhu* (Mason et al., 2021), but not all RPGs work in the same way (Hitchens and Drachen, 2008; Arjoranta, 2011), and other authors have experimented with new ideas. For instance, in *Everyone is John* (Villegas and Witt, 2019) the players act as an independent voice inside the mind of the single main character, John, instead of acting as independent characters that grow through experience. The GM is in charge of creating and describing the places where John wakes up, and also acting as a judge. As a totally different example, in *Alice is Missing* (Starke, 2020) the players act as friends of Alice, a missing teenager, who are reacting after the shocking news. However, in this game, the responsibilities of the GM are limited to drawing

Chapter 2. The State and Fate of Interactive and Improvisational Storytelling

cards from a deck and paying attention to a 90 minute timer that controls the pace of the game.

As can be seen, there are tons of RPG systems that differ in their rules and the expected themes. For the work I present in this thesis, we are mainly focused on modeling the classic gamemastering roles, such as the *Dungeon Master* (DM) from *Dungeons and Dragons*, the *Keeper* from *Call of Cthulhu*, or the *gamemaster* from Chaosium’s *Basic Roleplaying: Universal Game Engine* (Durall and Perrin, 2023), which has the particularity to offer the GM and players a diverse set of rules, so it can be used for roleplaying in almost any setting/theme (e.g. Fantasy, Lovecraftian horror or Space).

Summing up, since their inception, Role-playing Games have provided a framework for **Interactive** and **Improvisational** Storytelling. Through a dialogue where the GM and players participate, the players explore a world where all the living beings that live in it, as well as the events that happen in the environment, are typically controlled by the Game Master. Role-playing Games were so innovative that the influence of RPGs in the whole entertainment industry was huge (MacCallum-Stewart et al., 2018). As I will expand in the next section, video games were not left behind (Adams, 2009).

2.2 Computer Role-playing Games: The first gamemastering models

The 1974 release of D&D and its massive impact in popular culture, soon started an avalanche of computational efforts to recreate the tabletop experience as a computer game (Reed, 2023). Some of them were focused on the interaction with an automated GM using natural language, known as **Interactive Fiction**, while others were focused on programming all the rules to simplify the bookkeeping and let the player enjoy the strategic aspects of RPGs (Adams, 2009), known as **Computer Role-playing Games** (CRPGs). In this section I will show an example for each of them, hoping they will be illustrative enough to convey the conceptual notions of how these games work.

2.2.1 A natural language interface — Zork

The *Zork* games first released in 1977 — and all their contemporary text games as well — invited the players to change the way they interacted with a game: instead of using the keyboard to choose from a list of commands, they had to type what they wanted to do using one or two English words (Reed, 2023), as shown in Figure 2.6. This **user input** (or **player input**) was then processed by the computer to find what command(s) was the user referring to, and then execute it to update the fictional world state and move the game forward. In Computer science jargon, this process of analyzing a *string* according to a specific formalism is called *parsing*, and I will come back to it in Section 2.3.

The available commands were really a lot: get things, use things, move things, move the player, examine an item, examine a place, and many more (Infocom, 1984). In turn, each command was mapped to specific code instructions to update the world state, which the developers have **preprogrammed** in advance. As the same command (e.g. GET APPLE) could be used in different parts of the adventure, the parser also had to be aware of the state of the fictional world. This classic world-update process, based on parsing, is represented by the diagram in Figure 2.7: a parser processes the user input and determines which is the desired command among the possible ones,

2.2. Computer Role-playing Games: The first gamemastering models

```
>LIGHT LANTERN
The lamp is now on.
You are in a dark and damp cellar with a narrow passageway leading
east, and a crawlway to the south. On the west is the bottom of a steep
metal ramp which is unclimbable.
The trap door crashes shut, and you hear someone barring it.
>UP
The door is closed.
>EAST
You are in a small room with passages off in all directions. Bloodstains
and deep scratches (perhaps made by an axe) mar the walls.
A nasty-looking troll, brandishing a bloody axe, blocks all passages out of
the room.
The troll's axe barely misses your ear.
Your sword has begun to glow very brightly.
```

Figure 2.6: A snippet from the Zork I videogame. This image was directly taken from Reed (2023).

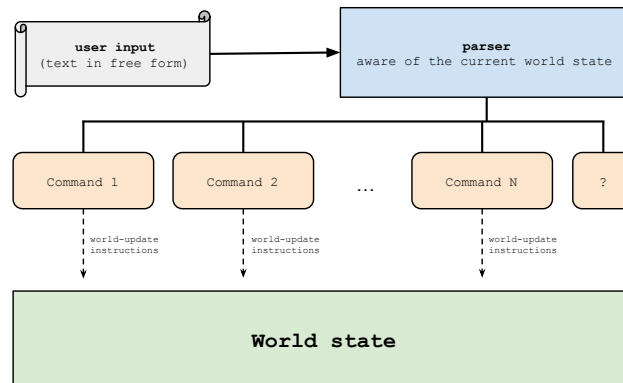


Figure 2.7: A diagram that shows how the main components interact in a classic world-update strategy used in text games. The narrative part of the system (i.e. how it describes the world to the player) is not represented here.

having the option to reject (represented by “?”) the user input if not recognized; then, the detected command is mapped to a sequence of instructions that update the world state accordingly, represented by a straight dashed arrow. In later versions of Zork, the parser could detect more than one command from a single user input (Reed, 2023).

To incorporate some narrative skills to the modeled GM beyond the world-update process, the code instructions executed after the parsing process could also contain previously written texts to describe something to the player e.g. “Taken” after a GET command, or “The door is closed” after trying to open a door that is locked (Reed, 2023). As expected, not all imaginable user inputs can be mapped to a finite set of commands, so this system also had previously written texts to indicate the player if its input was not a valid command, and were called “ZORK Complaints” (Infocom, 1984).

To summarize, Interactive Fiction systems like *Zork* were one of the first computational attempts to model a Game Master from a Tabletop Role-playing Game. They focused on three key abilities that a human Game Master should have:

- Analyzing a user input written in natural language (English, in that case), using *parsers*

Chapter 2. The State and Fate of Interactive and Improvisational Storytelling



Figure 2.8: Characters can move in a map and interact with the environment. This screenshot was taken from the “Baldur’s Gate 3 - PlayStation 5 Pre-Launch Trailer” video, uploaded by Larian Studios to YouTube, on September 2, 2023. Image shared according to “Larian Studios’ Fan Content Policy”.

- Properly updating the fictional world state after the user input, using pre-programmed logic for each possible action.
- Narrating scenes, describing places, objects and characters, using templates previously written by the designers.

This breakdown of the gamemastering skills was revolutionary, and they are the basis for the PAYADOR system I will present later in this document.

2.2.2 A huge closed world — Baldur’s Gate 3

Although **Interactive Fiction** games were hugely inspired in RPGs, the concept of **Computer Role-playing Games** (CRPGs) was historically used for games using a graphical representation of the world, instead of text descriptions (Reed, 2023), and strongly focused on combat and character growth (Adams, 2009).

As an example for this kind of games, I will pick *Baldur’s Gate 3*, a 2023 CRPG that implements the D&D 5th edition official ruleset. The game is really complex and can be analyzed focusing on its narrative design, graphics, combat system, and many other aspects. However, I want to focus on how much user agency this game, as an Interactive Narrative system, has to offer.

Baldur’s Gate 3 offers interactivity in two main modalities. The first one, shown in Figure 2.8, is an explorable world, in the form of a map, where the player can make the characters interact with the environment: moving around, opening containers, attacking NPCs, trading with other characters, and many more preprogrammed actions. As a personal comment, it is really impressive the great variety of actions to perform and how deep the interaction with the environment can be.

The second interactive modality, shown in Figure 2.9, is how the players can talk to NPCs. Instead of using free text, like in Interactive Fiction, here a branching narrative system is implemented. When the player participates in a dialogue, the decisions taken are saved in the game’s database to condition the way the story develops. Again, the amount of authored content this game has is really impressive, and many dialogue options are contemplated. Since some of these options usually present a major challenge for the player, *Baldur’s Gate 3* also introduces stochasticity: just like in the real

2.2. Computer Role-playing Games: The first gamemastering models

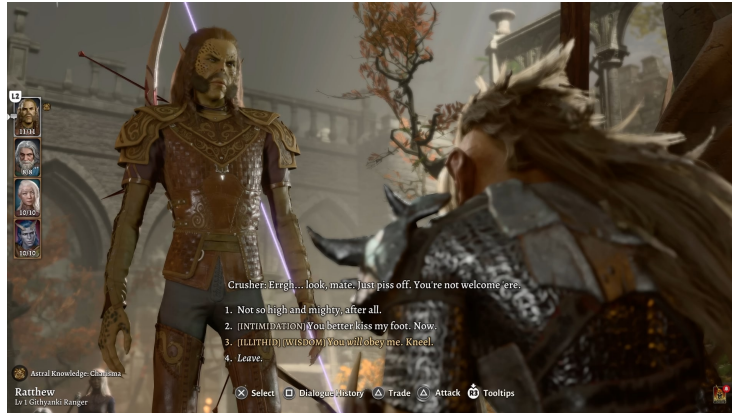


Figure 2.9: Dialogues are modeled as in classic branching narrative systems, where some options — like options 2 and 3 in this case — require dice rolls. This screenshot was taken from the “Baldur’s Gate 3 - PlayStation 5 Pre-Launch Trailer” video, uploaded by Larian Studios to YouTube, on September 2, 2023. Image shared according to “Larian Studios’ Fan Content Policy”.

D&D game, certain actions require dice rolls to determine whether they succeed or fail. However, regardless of this non-deterministic behavior, it is easy to note that this kind of system has the same level of user agency than the CYOA books, being the difference that here the available *nodes* are many more.

Narrative in Open worlds

When thinking about how much user agency a system offers, it is usual to wonder how agency can be measured (Harrell and Zhu, 2009), and how much it can grow when experimenting with improvisational systems (Swartjes, 2010). It can also raise the question of how much agency we have to narrate our own stories (Osler, 2024). When talking about games, it is usual to use the **Open world** concept to refer to those games where the player has few limitations in how to explore (in which order, or performing what actions) the fictional world, and Baldur’s Gate 3 fits in that category. However, since all those actions have been preprogrammed in advance, all those games are far less flexible than TTRPGs (Adams, 2009), which aim to offer the peak of interactivity in an Open World.

Many definitions can be considered when thinking about storytelling in open worlds. Some definitions make a focus on the limitations of the fictional world, and others on the limitations of the player’s choices. In order to guide the discussions included in this thesis, we will consider the definition of *Open-world Storytelling*¹ conveyed by Martin et al. (2016) in the following passage:

Humans have the ability to connect seemingly unrelated ideas together. If a computer is working together with a user to create a new story, the AI must be prepared to handle anything the human can think of. Even when given a scenario that appears constrained, people can — and will — produce the unexpected.

It is important to point out that this improvisational freedom is also constrained by common-sense (Grishakova, 2009): depending on the rules that govern the fictional

¹This concept has no relation to that used in the Databases research community, about evaluating truth conditions according to Open-world or Closed-world assumptions (Reiter, 1981).

Chapter 2. The State and Fate of Interactive and Improvisational Storytelling

world, there will always be some actions that are not valid there (Durall and Perrin, 2023). This has a strong relation to the main research challenge behind Interactive Narrative systems: how to adjust the player agency they offer so that the player's actions do not force the fictional world into an incoherent state (Benotti, 2006), a state in which the system cannot build off during following user-system interactions (Riedl and Bulitko, 2012). To solve this problem, Bates (1992) proposed the concept of *Drama manager*, an omniscient agent that oversees user-system interactions to ensure they meet some predefined standards for a *pleasing experience*. Later, Riedl et al. (2008) proposed an *Experience manager* as a more general concept to be used beyond *dramatic* interactive experiences (e.g. interactive systems for education). These concepts are closely related to the concept of a Game Master, who also actively monitors the player actions to ensure the consistency of the story and the fictional world: even when the players are experiencing *free will*, keeping the coherence of the fictional world is still a priority.

Increasing the size of a closed world

However, as I previously mentioned, *Baldur's Gate 3* does not allow free-text input and implements a non-deterministic branching narrative system. Therefore, this game is an interesting example of how free the players can feel when the authored content of a branching system is really big, even when they have not enough agency to propose anything they can think of. This is intrinsically related to what Martin et al. (2016) commented on closed-world systems that appear open.

Of course, this is not a mere coincidence. As of January 2025, the game's description on Steam² includes the following lines, some of which I underline:

(...) Baldur's Gate 3 gives you unprecedented freedom to explore, experiment, and interact with a thriving world filled with characters, dangers, and deceit (...) your actions define the adventure, but your choices define your legacy. You will be remembered. (...) Every choice you make drives your story forward, each decision leaving your mark on the world. Define your legacy, nurture relationships and create enemies, and solve problems your way. No two playthroughs will ever be the same. (...) Over 600 spells and actions offer near-limitless freedom of interactivity in a hand-crafted world where exploration is rewarded, and player agency defines the journey.

In the previous lines there is a noticeable effort of the publisher to emphasize that the game offers the player great agency, having the possibility to control how the story develops. They also state that *no two playthroughs will ever be the same*, what is achieved by combining the pre-authored content in many different ways (i.e. navigating through the branching narrative's graph). Judging by its universal claim³, both by critics and players, the strategy to preprogram a massive amount of content works really well. However, as the publisher says in the previous lines, it offers *near-limitless freedom*. So the question is: *How can we take a step forward?*

That question, like in a CYOA book, would take us straight to Section 2.4. But this thesis is not an interactive narrative system, so first, the next section will introduce some tools that are vital to model a Game Master that communicates with the rest of players using natural language.

²https://store.steampowered.com/app/1086940/Baldurs_Gate_3/

³<https://opencritic.com/game/9136/baldurs-gate-3>

2.3 Natural Language Processing

Historically, *formal languages* — like programming languages — have been the standard way to give instructions to a computer. These languages are designed by humans with the objective of being easily interpretable by the computer (i.e. machine readable), while reducing its ambiguity as much as possible (Scott, 2009). The languages we humans speak, such as Galician or Guarani, are not meant to be understood by a computer, but to communicate complex ideas between us instead. Since they are not designed with a specific goal in mind, and they evolve naturally according to ever-changing human needs, we call them *natural languages*. The field of **Natural Language Processing** (NLP) is concerned with creating computer programs that try to overcome these problems. The objective of an NLP model may be to automatically translate a text between two natural languages, extract the name of people mentioned in a document, write a narration of a scene from a fictional world, and others. Among them, there is a common problem that all NLP systems face, and it is the greater degree of ambiguity that natural languages have compared to formal languages, which, as I said, were designed to avoid it (Manning, 2015).

As I detailed in Section 2.1.3, the core mechanics of RPGs strongly rely on players’ imagination. The complex ideas living in their minds, such as the thoughts and intentions that are intrinsic to the identity of each character, cannot be shared with the gaming group using a finite set of unambiguous rules, like in the vast majority of board games (Tekinbas and Zimmerman, 2003). Instead, they are communicated using natural language, so the players’ ability to communicate those ideas in a rich dialogue between all of them, Game Master included, is also part of the game (Durall and Perrin, 2023).

As can be seen, to model a Game Master that can narrate and act as other characters, but also understand what the players want to do in the fictional world, we will need modules that can process natural language, the only way in which complex ideas can be communicated among the gaming group. In this section I will describe some concepts from the field of NLP that are at the core of the PAYADOR approach.

2.3.1 Rule-based models

The fact that some phenomena cannot be modeled with a fixed set of rules — like TTRPG dynamics — does not mean that rule-based strategies have no utility. They are very suitable for problems where a pattern analysis is enough, without risking the system’s performance due to problems arising from the lack of interpretability of modern neural-based NLP approaches (Danilevsky et al., 2020). Therefore, I will next describe and comment on two classic rule-based formalisms used in NLP.

Regular Expressions

Regular Expressions (RegEx) allow us to define formal languages according to some rules that describe their strings. More importantly for NLP, they also can be used to search patterns, which allow systems to take decisions according to an input in free-text form (Hopcroft et al., 2006). In fact, RegEx are used in a key module of PAYADOR that I will later describe in Chapter 4.

Since they are a well-documented concept, I will not go into details about their formal inductive definition or their expressive power. However, the following example tries to show the main available operators: concatenation, union ($|$), and Kleene star for repetitions ($*$).

$$L_1 = r(\text{Un camino muy laa*rgo|A very loo*ng road})$$

Chapter 2. The State and Fate of Interactive and Improvisational Storytelling

The formal language L_1 contains all the strings, both in Spanish and English, that have one or more *a* letters to say “largo” or one or more *o* to say “long”. That is $L_1 = \{\text{‘Un camino muy largo’}, \text{‘A very long road’}, \text{‘Un camino muy laargo’}, \text{‘A very loong road’}, \text{‘Un camino muy laaargo’}, \text{‘A very looong road’}, \dots\}$.

In modern implementations, like in the Python3 `re` module⁴, there are other operators available that were built combining the basic operators from the theoretical definition.

Context-free Grammars

Context-free Grammars (CFGs) also have operators that combine patterns, but they have even more expressive power, hence specify languages that RegEx cannot. This extra power comes from a key difference: the recursive structures they can model. This allows to use them to build a *parse tree*, a tree representation of the structure recognized by their rules, which has a lot of information that can be used to implement text-based processes.

That idea was so powerful that CFGs have been widely used in compilers for programming languages (Hopcroft et al., 2006), and also were the main engine in NLP systems for a long time, both for analyzing inputs in free-form text and for generating language (Chomsky, 2002; Jurafsky and Martin, 2008). In fact, Interactive Fiction games are also called *Parser games* since they implemented parsers to identify the actions the player wanted to do and update the fictional world accordingly, as represented by Figure 2.7. The first games used parsers that analyzed simple VERB-OBJECT syntactic structures (e.g. to recognize GET APPLE), while those in later games like Zork analyzed more complex ones (Reed, 2023).

Although the PAYADOR system does not have any modules that use CFGs, this section included them to remark an important difference: in contrast to the classic world-update strategy where parsers were used, in the PAYADOR approach we tried to leave out the parsing step to rethink the whole process, as I will further discuss in Chapter 4. Instead, PAYADOR relies on Large Language Models, the state-of-the-art formalism in Natural Language Processing, that I will describe in the next section.

2.3.2 Large Language Models

Back in 2017, the *Transformer* architecture (Vaswani et al., 2017) was proposed as a way of dealing with texts of variable length (i.e. number of tokens) and having a flexible strategy to model its meaning, based on the efficient *word embeddings* representation proposed by Mikolov et al. (2013). Although the Transformer was originally used for Machine Translation, this architecture soon allowed the design of modern *language models* that changed the whole NLP field (Min et al., 2023).

One of those architectures for language models was BERT (Devlin et al., 2019), that used the *encoder* part of the Transformer architecture to learn word representations. Following a strategy called *masked language modeling*, the neural network is trained to predict which token is likely to fill the masked ones in a sequence (e.g. which token can replace the [MASK] in “A very [MASK] road”). The resulting vectors have different representations for each word, depending on the context it appears in. That is the reason why they are known as *contextual word embeddings*, while those proposed by Mikolov et al. (2013) are known as *static word embeddings*, as they represent each word with only one vector.

The other remarkable architecture is GPT (Radford et al., 2018), that brought back the 20th-Century idea of modeling how language works by calculating the prob-

⁴<https://docs.python.org/3/library/re.html>

2.3. Natural Language Processing

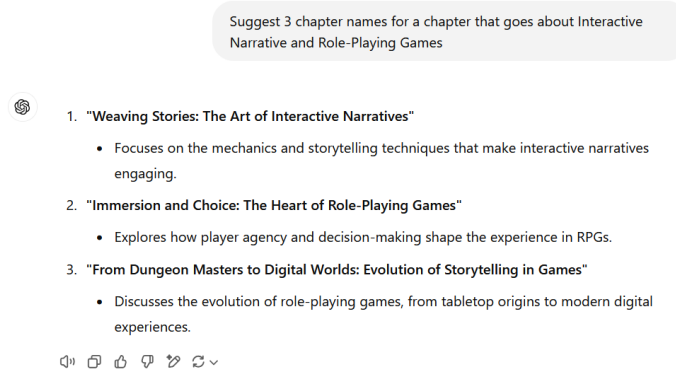


Figure 2.10: A simple example of conditional generation using ChatGPT on January 23, 2025.

ability of a sequence of words. In the past, n-gram models were used to model such probabilities (Brown et al., 1992), but GPT used the *decoder* part of the Transformer architecture instead. After the original version, OpenAI worked on GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023). These last models shaped the modern era of *conditional generation*, a task consisting on generating text conditioned on other text (Maynez et al., 2023). The initial text needed to start generating new text is called a *prompt*, and usually consists of instructions for what we want the model to generate. In Figure 2.10 there is an example of generating text using ChatGPT⁵, where the *prompt* is “Suggest 3 chapter names for a chapter that goes about Interactive Narrative and Role-Playing Games”.

The strategy to generate new text is generally always the same: generate a token according to the probability that it can continue a previous sequence of tokens, and then consider all of them to generate another token, and so on. This method, called *autoregressive generation* (Maynez et al., 2023; Zhou et al., 2024), worked so well that researchers started to try different strategies to get a better performance. One of those possibilities is to adjust the model to be specialized in a specific domain or task. This process, called *fine-tuning*, consists of continuing to train the model on data that better represent that domain or task (Zhou et al., 2024). Another possibility is to explore diverse *prompting methods* (Liu et al., 2023) (i.e. different ways of writing the prompts). In this work we focused on two of them: *zero-shot learning* and *few-shot learning*, originally proposed in the GPT-3 paper (Radford et al., 2019).

Zero-shot learning refers to the strategy used in Figure 2.10, where the prompt consists of instructions for the desired generated text. Instead, when the prompt contains examples, the method is called *Few-shot learning*. An example is shown in Figure 2.11, where ChatGPT answers right after being prompted to translate from English to Rioplatense Spanish. As can be seen, the prompt does not include any instructions, and it consists of just an example and an empty line for the desired translation. However, *few-shot* prompts can also include instructions, such as introducing the problem to be solved, or adding comments to the included examples. As Chapter 4 will later detail, PAYADOR uses *few-shot* prompts to perform the *world-update* process.

Beyond *zero-shot* and *few-shot* learning there are other relevant prompting techniques. *Chain-of-thought* prompting (Wei et al., 2022) instructs the model to generate the result after a step-by-step reasoning. The authors empirically show that

⁵<https://chatgpt.com>

Chapter 2. The State and Fate of Interactive and Improvisational Storytelling

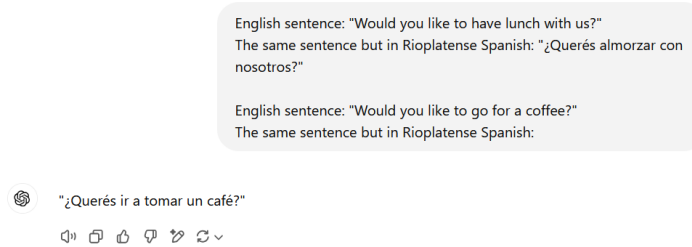


Figure 2.11: A translation from English to Spanish by prompting ChatGPT using few-shot learning. January 23, 2025.

these prompting strategy improves the LLM performance on arithmetic and reasoning tasks. *Self-consistency* (Wang et al., 2023) uses the main idea behind *chain-of-thought* prompting, but trying to generate different thought processes for the same problem and then voting for the most frequent result (e.g. if 3 thought processes suggest that the result is “2” and only one suggests that is “5”, then the chosen option is “2”). *Tree-of-thought* (Yao et al., 2023) also uses the main idea behind *chain-of-thought* prompting, but this time trying to explore different reasoning paths. This opens the possibility of framing problems as a search over a tree, where each node is a possible intermediate *thought* towards the final solution. Finally, the recent *Derivation* prompting (Sastre et al., 2025) uses logic derivations to search the answer space in the context of Question Answering. The method tries to derive conclusions from initial hypotheses, applying rules that help the LLM to do a step-by-step reflection during the reasoning process. Among all these techniques, we chose *zero-shot* and *few-shot* learning since they do not need a great amount of data to build the prompts, but also because they usually need less resources than the rest of methods (i.e. usually just one inference is enough to get a result).

These models based on the Transformer architecture, which from now on I will refer to as *Large Language Models* (LLMs), quickly changed the NLP field, achieving state-of-the-art performance in many tasks and across many domains (Min et al., 2023). In the next section we will see that Improvisational Storytelling is among the many problems that are far from being solved by these type of models (Ignat et al., 2024), even when they are *fine-tuned* on specific data.

2.4 NLP-powered models of Role-playing Games

So we want to model *Game Masters* who communicate with the rest of the players using *natural language*. The players can freely act in an open world, so the GM surely has to improvise after actions that were not contemplated as a possibility, but they are completely *valid* (coherent with the story and state of the world). Naturally, NLP tools are mandatory to model these games with conversational nature.

Although in the past there were some attempts at modeling the complexity of TTRPGs relying on complex logic systems or sets of rules (Peinado and Gervás, 2004, 2007; Tychsen et al., 2009; da Rocha Franco et al., 2015; Franco et al., 2018), it was not until the era of neural language models that there was an explosion in the number of published works. However, as I will show in this section, that does not imply that logic systems are no longer used. Those systems that combine neural models with rule-based approaches on structured data (e.g. graphs, databases), are usually called

2.4. NLP-powered models of Role-playing Games

neuro-symbolic.

In the last section of this chapter I will present recent work on modeling some aspects of TTRPGs, GMs included. This section will not only include works that explicitly use NLP models, but also related works that were published during the era of LLMs, hence were conceived having in mind the new possibilities we have. Also, it only focus on works that explicitly try to model concepts related to TTRPGs, leaving out CRPGs or other video game genres. The surveys of Gallotta et al. (2024) and Sweetser (2024) cover many works on using LLMs in video games, including those with a big influence of TTRPGs.

2.4.1 Assistants

In order to understand what kind of tools could be built to assist GMs to run game sessions, there are some works on doing interviews with GMs and Players and building virtual assistants.

Acharya et al. (2021a), Acharya et al. (2021b) and Tang et al. (2023) interviewed GMs to understand how technology can assist them. All of them reported the results following a systematic methodology and including excerpts from the interviews, what make the agreements and discrepancies between the participants easy to understand.

Santiago III et al. (2023) performed preliminary experiments with GPT-3 as a companion for D&D. The companion, named *Avalon*, gives suggestions to the Dungeon Master in specific scenarios that are described in the prompt. Acharya et al. (2023) proposed *Shoelace*, a system to assist GMs using a Graph representation and Prolog. Later, Kelly et al. (2023) extended it to include dialogue suggestions for NPCs. Ito and Okamoto (2024) explored the possibility of building automated NPCs to assist human GMs and reduce the amount of work.

Zhu et al. (2023b) proposed *CALYPSO*, an LLM-based system that helps Dungeon Masters to run a D&D session. It offers a chatbot that can summarize information from the D&D official rules, help with questions about the ongoing session, and help DMs to brainstorm ideas. They evaluated the system on a Discord server that serves as a *play-by-post* gaming server. In these cases, the game is played through written utterances, not spoken dialogue, which facilitates the use of NLP tools, both for processing them and for calling an automated assistant to apply some rules.

2.4.2 Datasets and Environments

There are some datasets of dialogues taken from real TTRPG sessions. For instance, Louis and Sutton (2018) collected a dataset from the Roleplayer Guild *play-by-post* forum⁶. The dataset contains more than 56,576 posts in English. Rameshkumar and Bailey (2020) created a dataset consisting of almost 400 thousand dialogue turns of game sessions in English. These dialogue turns are transcribed from the popular *Critical Role* show, where voice actors play Dungeons & Dragons (see Figure 2.12). Last but not least, Zhu et al. (2023a) presented FIREBALL, a dataset of 8 million utterances from 25 thousand D&D sessions played in English on Discord. All these sessions were played using Avrae, a Discord bot that facilitates D&D gameplay online, allowing them to later fine-tune an LLM to generate some of its commands.

Other datasets are focused on the rules, official examples, or other textual data related to TTRPGs. Peiris and de Silva (2022) present the Forgotten Realms Wiki dataset. The *Forgotten Realms* is one of the most popular settings for D&D, and the wiki with that name⁷ collects and organizes data about it. The dataset is in English,

⁶<https://www.roleplayerguild.com/>

⁷<https://forgottenrealms.fandom.com>



Figure 2.12: A screenshot taken from the “Bathhouses and Bastions — Critical Role — Campaign 2, Episode 90” video, uploaded by Critical Role to YouTube, on January 13, 2020.

and its raw version contains about 9 million tokens. Wu et al. (2024) published a dataset of several documents in Chinese related to Jubensha, a type of RPG where players try to find who is guilty for the murder they are investigating.

Finally, there are some datasets that work as environments to perform experiments on text games (e.g. Interactive Fiction). For instance, LIGHT (Urbanek et al., 2019) contains thousands of Objects and Characters, hundreds of Locations, thousands of dialogue lines, and many possible actions to interact with the fictional world. The LIGHT environment was later extended by Ammanabrolu et al. (2021) to include *quests* (i.e. objectives). TextWorld (Côté et al., 2019) is a library for Python with algorithms to quickly set up different environments. Jericho (Hausknecht et al., 2020) offers a Python environment to connect automated players to previously existent Interactive Fiction games. All these datasets/environments have their components (locations, characters, objects, quests, actions, dialogues, and so on) in English. Based on the possibilities that these environment offer, there is also a lot of work in training Reinforcement Learning agents to play text games, and in general interact with worlds described using natural language, what can be seen as modeling a TTRPG player. Since they are out of the focus of this section, they are not included here, but the surveys of Osborne et al. (2022) and Jansen (2022) cover many of them. There are also numerous efforts in making LLMs act as agents with motivations and personality — like characters — what is generally called *role-playing*, and the survey of Chen et al. (2024) covers many of them.

2.4.3 Generation of content

Some responsibilities of a GM demands creating content for the players. This can be done before the session, when preparing it, or during the session, where improvisation is key. Given the many restrictions that a GM has, this constitutes an interesting mechanism to be modeled (Guzdial et al., 2020).

The most typical thing a GM has to improvise are NPCs answers when the players interact with them. Huang (2024) used LLMs to offer the player interactive dialogues for the “Alone Against the Flames” module of the Call of Cthulhu TTRPG. This module consists of a single-player adventure to introduce new players in the game mechanics, and works similar to a branching narrative, where the player reads the out-

2.4. NLP-powered models of Role-playing Games

comes of a limited set of options. More generally, Zhong (2023) proposed to generate scripts based on D&D data, taking into consideration different settings and characters.

It is also useful if the GM can create custom content, mechanics or rules that better fit their fictional worlds (Durall and Perrin, 2023). With that in mind, Newman and Liu (2022) and Musacchio et al. (2024) built datasets of official and fan-made spells for D&D. Then, they used them to generate spells that could actually work in real D&D sessions.

Another important creative task of a GM is to improvise locations, characters and objects, how they look like and how they behave. For neuro-symbolic approaches, it becomes critical to have models that can generate narrations from structured representations, and also data that fit in those structures. Callison-Burch et al. (2024) presented DAGGER, a dataset generated by GPT-4 that expands LIGHT. They also explore the generation of narratives to describe structured fictional worlds, and vice versa. Chambers et al. (2024) presented BERALL, a neuro-symbolic system that combines LLMs and structured representations to generate playable worlds for Interactive Fiction games.

2.4.4 Other gamemastering mechanisms

There are other works that focus on mechanisms that GMs use to keep the players interested and engaged.

Xi et al. (2021) modeled a GM that sets an objective for the player to achieve. The NLP pipeline processes each dialogue utterance to decide if the player has completed (or not) the objective. Zhou et al. (2023) focused on modeling the guidance that Game Masters usually give to the player towards a specific objective. They do this on the data collected by Callison-Burch et al. (2022), that I will describe below. Carlander et al. (2024) analyzed the effect of how different prompts affect the performance of an LLM on detecting if a player is speaking in-character or out-of character. This decision is based on how the player writes, but also on the alignment of the character, a concept of D&D.

2.4.5 General gamemastering models

The rest of works I will include in this section try to model the general TTRPG experience, directly or indirectly modeling a GM.

Shuster et al. (2021) proposed the interesting approach of using the LIGHT dataset to model a GM, let people play with it and use those dialogues utterances as data to train language agents. They are, in fact, modeling how humans interact in a Role-playing game environment. Callison-Burch et al. (2022) created a dataset from the official D&D play-by-post forum⁸. Then, they use it to fine-tune a language model in order to improve its performance on two tasks: game state tracking and response generation based on a specific game state. The resulting model — a model of the whole D&D game — can be used as a player or the Dungeon Master, showing good performance on the narrative generation task. However, one of their main conclusions is that additional machinery would be needed to track the game state beyond a Language Model. Song et al. (2024) modeled a GM for the “Jim Henson’s Labyrinth: The Adventure Game” TTRPG. They follow a neuro-symbolic approach, where the GM is grounded to a structured representation, which is updated via LLM function calls. Finally, Zhao et al. (2024) built a system based on LLMs where the players can interact with the environment and NPCs in detective and adventure stories.

⁸<https://www.dndbeyond.com/forums/d-d-beyond-general/play-by-post>

Chapter 2. The State and Fate of Interactive and Improvisational Storytelling

After the popularization of ChatGPT, some researchers performed preliminary experiments using it as an out-of-the-box gamemastering model. Saito et al. (2023) reported an experiment in which children used different GPT models to design serious RPGs (i.e. games with an additional purpose beyond entertainment) with themes that they found interesting. They later played that game with the LLM acting as the Game Master. Triyason (2023) designed prompts to ask ChatGPT to do something similar: act as a Dungeon Master, the Game Master of the Dungeons & Dragons game. After performing experiments with players and experimented human Dungeon Masters, they were asked to score and evaluate their experience. You et al. (2024) also designed prompts with the same purpose, but they tried with three different versions. Then, they performed experiments to analyze which were the emotional reactions of the players to the different versions of the modeled Dungeon Masters.

2.4.6 AI Dungeon

One of the most famous general models of the Player-GM interaction is *AI Dungeon*⁹. In this game, the GM starts describing a situation — whose details are usually included in a prompt previously written by a human — to communicate the player where the character is, what the characters were doing before starting to play, and so on. Then, the player can speak to the GM in free-text form, which triggers several prompts for the GM to keep narrating the outcomes. According to the information included in the *AI Dungeon Guidebook*¹⁰ the system is not grounded to a structured representation nor uses logical engines to manage the fictional world. They do have an extra structure called “Plot essentials”¹¹, where the player can include relevant facts of the plot to be remembered. However, everything is limited by the token length, which indicates that — in the end — everything is LLM-based.

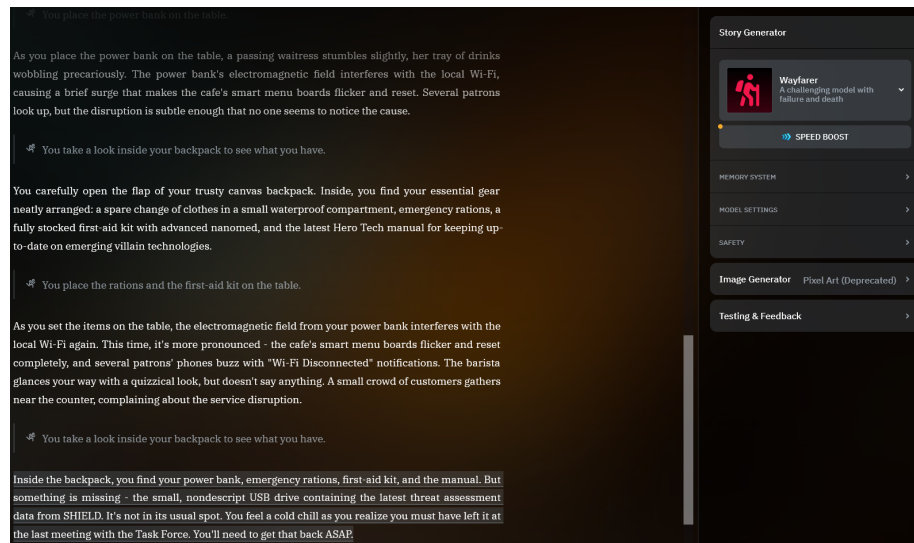


Figure 2.13: Screenshot of an AI Dungeon gameplay, in the *Superhero* genre. The modeled GM fails in keeping track of the location of objects. Image taken on January 22, 2025.

⁹<https://aidungeon.com/>

¹⁰<https://help.aidungeon.com/>. Last visited January 26, 2025.

¹¹<https://help.aidungeon.com/faq/plot-essentials>. Last visited January 26, 2025.

2.5. Closing remarks and key points

As Callison-Burch et al. suggested back in 2022, LLMs seemed not to be enough to track the full game. Although present-day LLMs have stronger capabilities, they are still far from solving this issue. For instance, Yong and Mitchell (2023) performed an observational study of 10 participants playing AI Dungeon. Among the several things they analyze, they try to understand how the lack of story coherence or world inconsistency affect the engagement and immersion of the players.

We tried the game using one of the last fine-tuned models available, Wayfarer¹², published at the end of 2024. After moving some objects that were included in the inventory, we asked the model what objects my character had in it. The fictional world resulted to be inconsistent, as shown in Figure 2.13.

2.5 Closing remarks and key points

In this chapter I defined the main concepts that are needed to frame the work presented in this document. From the Choose-your-own-adventure book series to Baldur's Gate 3, interactivity opens many possibilities for storytelling. However, TTRPGs still are the only ones offering the highest degree of player agency. LLMs provided a new opportunity to work on Open-world Interactive Storytelling, but the problem of having a general Gamemastering model, explainable and reliable, is far from being solved.

The key points of this chapter are:

- **Interactive Storytelling** or **Interactive Narrative** are those systems in which the reader (or user) has some degree of control on how the story develops.
- **Tabletop Role-playing Games** (TTRPG) are class of games — played through dialogue — in which the players act as characters in a fictional world. Typically, the game is orchestrated by a **Game Master** (GM), who communicates with the rest of players.
- **Non-playable character** (NPC) is a character that is not controlled by a player and it is controlled by the GM instead.
- **Interactive fiction** (IF) is the name given to a subset of digital Interactive Storytelling systems that modeled a TTRPG Game Master focusing on two of their basic skills: updating the fictional world state after a **user input** (or *player input*), and describing the scene. To interact with these systems, the player uses text in free form, that is later *parsed* by the computer.
- **Computer Role-playing Games** (CRPG) are a different approach to adapt the original TTRPG experience. They are mainly focused on strategy, character management and character development.
- **Player agency** (or user agency) is the amount of freedom the player has to change the fictional world state, or — in general — the course of the story narrated by the system.
- **Large Language Models** (LLMs) are a type of model that can predict the probability of a token happening after (or in the middle of) a sentence. Nowadays they are mainly used for generating text following the *autoregressive generation* technique. This process is based on prompts, a piece of text that is written to give instructions about what we want to be generated.
- **Prompting methods** are different techniques to write effective prompts. In this work we used *zero-shot learning* (the prompt includes only instructions) and *few-shot learning* (the prompt includes examples).

¹²<https://huggingface.co/LatitudeGames/Wayfarer-12B>

Chapter 2. The State and Fate of Interactive and Improvisational Storytelling

- **Neuro-symbolic** approaches are those where symbolic (e.g. rule-based models) and neural models (e.g. LLMs) are combined to take advantage of the best characteristics of both approaches.
- Many TTRPG models rely entirely on text to directly mimic the dialogue experience. Others, follow a mixed approach, adding annotations to the text to be processed by an LLM. There are also other neuro-symbolic approaches where the LLM calls functions that update some structured representation of the game.

As Puchal and López (2021) analyzed, the GM functions are diverse, complex and numerous. That is the reason why many of the works cited here focus on specific functions, like the emotional aspect of narrative, the consistency of the world, the guidance towards objectives, and so on. The available TTRPGs, their themes and settings, are also diverse. Although AI Dungeon offers the possibility of playing sessions with different settings, such as *Cyberpunk*, *Science Fiction* or *Superhero*, many of the works cited here build datasets or try to explicitly model concepts from the Dungeons & Dragons game, with a fantastic theme by design. As I will show in the next Chapter 3, fantasy is usually the chosen setting by an LLM when prompted to act as a GM in a TTRPG. This is very reasonable, since D&D is considered to be the first TTRPG to be published, and the most popular one. Therefore, the possibilities of gathering data for this domain are substantially reduced if another game is chosen.

While the vast majority of efforts cited here (e.g. fine-tuned LLMs, datasets and environments) are for D&D and English, our work is focused on proposing methods to model general gamemastering abilities (i.e. trying to be as system-agnostic as possible) that can be used both for English and Spanish. In particular, our work is mainly focused on modeling how GMs describe the fictional world using natural language, how they embrace the players' creativity when acting in that world, and how they consistently update the state of that world after the players' actions. I will come back to this in the upcoming chapters.

Finally, during the writing of this thesis, Shyne and Cooper (2025) published a scoping review covering many of the works cited in this chapter, also including the **PAYADOR** approach I will later present in Chapter 4.

Chapter 3

In Quest for the Neuro-symbolic Truth

But he believed
In a different truth
The heliocentric one
Proposed by Kopernikus
A new age has begun

Haggard
The Observer

Where to start...? That was the first thing I thought when faced with the idea of modeling a GM for the first time. Many possibilities, many subproblems to work on. Suddenly, after years of a strong debate in the NLP community whether neural models could really solve everything, OpenAI published ChatGPT, deepening the aforementioned debate (Ignat et al., 2024). Naturally, after some weeks thinking about the matter, we were wondering the same: *can ChatGPT act as a Game Master of a TTRPG? Are there any gamemastering problems that remain unresolved?*.

We performed preliminary experiments prompting ChatGPT to act as a GM, similar to those included in Section 2.4, and it quickly became evident that those problems reported by Callison-Burch et al. (2022) still happened: a neural-based language model is not enough to keep track of the state of the world. Coherence problems really occurred very often. However, the experiments we did were somewhat informal: we needed a systematic methodology to test gamemastering models. But the evaluation of creative systems has always been problematic due to lack of consensus on different aspects (Jordanous, 2012; Colton and Wiggins, 2012), and systems modeling TTRPGs are no exception. *How can we evaluate such models?* was a recurrent question for several weeks.

In this chapter I will present the path we followed to end up answering that question and, as an unexpected but welcome consequence, conceiving the main idea for PAYADOR.

3.1 Taking subjectivity out of the equation

One way to evaluate such models is to ask people to use them and analyze their playthroughs and the comments they want to do afterwards. Since we are modeling improvisational storytellers, that methodology would imply many subjective considerations and many variants of several phenomena, such as if the player should have done something else, or what could have happened in case of taking another decision. Although that is a classic methodology, it has many variables to consider in order to make a final judgment about what those models can — and cannot — do. Therefore, we wanted to explore something different, something that could complement that full evaluation with humans, something that can always be tested in an isolated way, even if the model is capable of running an open-world improvisational session and the player agency is at a maximum. We wanted to identify something that could be tested unitarily, just like unit tests work in software testing, without involving — too many — subjective considerations. What we needed to find were some characteristics that all gamemastering models should have, regardless of the technology used to build them. The name of this section is inspired in the core idea of the evaluation strategy I will describe next: given a domain, identify some key concepts that can be isolated, and think how to constrain the problem to evaluate them, transforming an open-domain evaluation into a closed-domain one.

According to what I have already discussed, a classic Game Master — like in *Basic Roleplaying: Universal Game Engine* (Durall and Perrin, 2023) — must give the players a feeling that every change in the world could be explained: things and people move, and events happen for a reason — even if the characters cannot understand the phenomena yet. In text games like *Zork*, those changes in the world state were performed in an exact and calculated way. Naturally, if every possibility has been contemplated by the designers (e.g. open a specific door with a specific key), and preprogrammed, the consistency is guaranteed. However, in an open-world system, it is not possible to preprogram every player’s choice at that level.

Among all the possible actions, there are some which are **irreversible** (e.g. it is not possible to get the original unbroken bottle after breaking it, except in some magical settings). Other actions are **reversible** (e.g. if *Aida* goes from the living room to the kitchen, she will be able to go back to the living room, except in specific narrative situations), what implies that if the reverse action is taken after the original one, the world state should remain similar (e.g. “go from the living room to the kitchen” + “go from the kitchen to the living room” = the character is in the living room). Since this intrinsic characteristic of reversible actions may help human evaluators to focus only on checking the outcomes of atomic changes in the world (without having many valid possibilities to consider), these are the kind of actions we chose to evaluate in narratives co-created with these open-world gamemastering models.

Therefore, we propose three *test categories* that can be used to create individual *tests*¹ to evaluate LLMs. The first two categories correspond to two classic *reversible* actions that were also modeled in *Zork*-like games: *object movement* and *player movement*. Besides those based on reversible actions, we also decided to include a third *test category* based on the most basic GM-Player interaction, mentioned in many TTRPG rule books: the GM narrates a situation, the players say what they want to do there, and the GM narrates the outcomes (Durall and Perrin, 2023). These *test categories* are not intended to be exhaustive, or to check a sufficient condition for gamemastering

¹In other words, these are like *templates* to instantiate different individual *tests*. For instance, the *Item tracking* category can be used to create a test to track items in a kitchen, another one to track items in a railcar, and so on.

3.1. Taking subjectivity out of the equation

models. Instead, we designed them to check some necessary conditions that should be present in any gamemastering model. I will describe all of them next.

3.1.1 Item tracking

The first test category, *Item tracking*, tries to determine if the model leaves the fictional world consistent after moving an object from the *inventory* to another place, and vice-versa. The idea is simple: list the items in the inventory, leave one of them somewhere else, and then check if the changes were made. It consists of the following steps:

1. *Player*: “What is in my backpack?”
2. *GM*: $Object_1, Object_2, Object_3, Object_4, \dots$
3. *Player*: “I take $Object_K$ and leave it on *another location*”
4. *GM*: *Narrates outcome*
5. *Player*: “What is in my backpack?”
6. *GM*: $Object_1, Object_2, Object_3, \dots$
7. *Player*: “What is on *another location* ?”
8. *GM*: $Object_A, Object_B, \dots$

If in step #6, $Object_K$ is still in the backpack, then the test is failed. The step #7 is not always needed, but it is useful when the model updates the inventory but does not update the available items in the location where the action takes place. Although in the previous steps there is only one object mentioned ($Object_K$), this test category can also be used to test situations when the player wants to take more than one object from the backpack, as in the example shown in Table 3.1.

This test category can be used to test any kind of model, LLM-based or not, since it is based on comparing lists of objects. In fact, this is the test used to check AI Dungeon’s world consistency in Section 2.4 (Figure 2.13).

| | |
|---------------|---|
| GM | (...) Aside from the mysterious key, your backpack also contains the usual adventurer’s equipment: a sturdy rope, a grappling hook, a water flask, some dried rations, a small first-aid kit, a compass, a map of the surrounding area, and a trusty dagger at your side. |
| Player | I take the rope, the map and the key and I leave them on the floor |
| GM | You carefully set the rope, map, and key down on the floor, making sure they are within reach in case you need them later. (...) |
| Player | What’s in my backpack? |
| GM | As you take a closer look at your backpack, you notice that it contains a few essential items that any adventurer would find useful: A sturdy rope: (...) this rope is perfect for climbing or rappelling down steep surfaces. (...) A map of the surrounding area: this map shows the nearby landmarks, towns, and topography of the region, allowing you to plan your route. (...) |
| Result | Failed: The rope and the map are still inside the backpack |

Table 3.1: An example of an *Item tracking* test, using ChatGPT as a gamemastering model.

Chapter 3. In Quest for the Neuro-symbolic Truth

3.1.2 Map consistency

In order to check the spatial consistency of the world, particularly focusing on the reachable places, the *Map consistency* test category is analogous to the *Item tracking* one:

1. *GM*: You are in $Location_1$ and, from here, you can access $Location_2, Location_3, \dots$
2. *Player*: “I go to $Location_K$ ”
3. *GM*: You are in $Location_K$ and, from here, you can access $Location_1, Location_A, \dots$
4. *Player*: “I go to $Location_1$ ”
5. *GM*: You are in $Location_1$ and, from here, you can access $Location_2, Location_3, \dots$

The list in step #3 should include the starting location from step #1 and the list in step #5 should be identical to the one in step #1, unless there is a narrative reason for that (e.g. The entrance to the cave has collapsed). If the locations are not consistently listed, the model fails the test. Just like the *Item tracking* category, this one can also be used to evaluate any gamemastering model since it only needs to compare lists, as shown in Table 3.2.

| | |
|---------------|--|
| GM | (...) Current place: You find yourself standing at the entrance of a dark and damp cave. The entrance is surrounded by tall trees, and the sound of running water can be heard in the distance. Accessible places: Inside the cave, Back to the forest, Along the river |
| Player | I want to explore the cave |
| GM | (...) Current place: Inside the cave Accessible places: The cave entrance, Deeper into the cave |
| Player | I want to go outside |
| GM | (...) Current place: Outside the cave Accessible places: Inside the cave, Along the river, Deeper into the forest |
| Result | Passed: The accessible places from the cave entrance are coherent. |

Table 3.2: An example of a *Map consistency* test using ChatGPT as a gamemastering model.

The *Item tracking* and *Map consistency* test categories may resemble the bAbI tasks proposed by Weston et al. (2016). Those tasks aim to provide a framework to evaluate Question Answering systems that need to reason over a set of facts (e.g. “John picked up the apple. John went to the office. John went to the kitchen. John dropped the apple. Where was the apple before the kitchen?”). However, the bAbI tasks were designed to have a single word or a list of words as expected outputs in a single Question-Answering interaction (e.g. “office”, for the previous example), what naturally is not the case for TTRPGs or Open-world Improvisational Storytelling in general. The interactive nature of TTRPGs really makes a difference in how the evaluator judges the GM model, as every player input will condition what the model will narrate next. As I mentioned earlier in this section, our test categories try to deal with problems naturally arising from that iterative co-creative process, trying to turn an open-domain evaluation into a closed-domain one.

3.1.3 GM-P-GM

The final test category we propose is inspired in the most basic interaction between the Game Master and the player (Durall and Perrin, 2023), hence we call it the *GM-P-GM* test. An example can be seen in Table 3.3 and the core idea is summarized in the following lines:

3.2. Experiments with three LLMs

- *GM*: Narrates a **situation** to solve in some **context**.
- *Player*: Describes the **actions** to overcome that **situation**.
- *GM*: Validates if those **actions** are feasible for that **context**, and next narrates the **outcomes**.

We proposed this test category to have a well-defined framework to evaluate LLMs as out-of-the-box Game Masters and, specifically, the commonsense reasoning abilities of the models, which are one of the *Achilles heel* of LLMs (Toroghi et al., 2024). This test category is the most subjective of the three proposed, and can be used in two ways:

1. The evaluator writes both the **situation** and the **context**, and also the **actions** taken by the player.
2. Having a world state already represented, check if the model can detect an invalid action.

As I will show next, our experiments followed the first approach as we wanted to explore complex narrative situations. The second approach is mainly suitable for systems that already have an initialized world state, hence the commonsense-reasoning evaluation arises naturally as a consequence of the user inputs. Natural narrative situations faced during real gameplay, such as the experiments I will report in Chapter 5, also belong to this category.

| | |
|---------------|---|
| GM | You enter the building to look for the lost suitcase. However, the door is covered in flames. To your right you see a fire extinguisher, which seems to be ready to use. You can't use anything from your backpack because you can hardly move your arms. What will you do? |
| Player | I take a bucket of water from my backpack and put out the fire |
| GM | As the player grabs the bucket of water from their backpack, they realize it's not nearly enough to put out the flames completely. However, they quickly pour the water onto the fire, creating a small opening in the flames (...) |
| Result | Failed: The character cannot grab the bucket of water from their backpack. |

Table 3.3: An example of a *GM-P-GM* test, using ChatGPT as a gamemastering model.

3.2 Experiments with three LLMs

We used the three test categories to evaluate the out-of-the-box gamemastering skills of three LLMs available at that time: ChatGPT², Bard³ — now Gemini — and the LLaMA-based OpenAssistant chatbot⁴. In order to do so, for each LLM we ran five tests corresponding to each test category (5 *Item tracking* tests, 5 *Map consistency* tests and 5 *GM-P-GM* tests), both in English and Spanish.

To create the five *GM-P-GM* tests, we wrote five different prompts with contradictory narrative situations (i.e. the player responds to the GM's challenge with an illogical approach), and to create the five *Item tracking* tests we wrote prompts that propose different starting locations (room, bathroom, kitchen and basement). For the

²ChatGPT March 14 2023 Version. "Free Research Preview"

³"Bard Experimental". Accessed on July 17th, 2023.

⁴"OpenAssistant/oasst-sft-6-llama-30b". Accessed on July 14th, 2023.

Chapter 3. In Quest for the Neuro-symbolic Truth

Map consistency tests we created a single prompt to use across the five different tests. We decided this in order to let the LLM create a world from scratch (i.e. the LLM creates a starting location and some reachable locations from there), what we think is an interesting characteristic of these tests.

The 5 *GM-P-GM* prompts, the 5 *Item tracking* prompts, and the single *Map consistency* prompt are available in Appendix A, where we include all the exact **strings** fed into the LLM to initiate the experiments in English. The prompts to run the tests in Spanish are just translations of those prompts. Table 3.4 shows the overall structure of the prompts, beyond the detailed versions that are available in the aforementioned Appendix A.

In summary, we evaluated 3 LLMs across 5 tests for each of the 3 *categories* (i.e. 5 runs for each *category*). As we did this both in English and Spanish, this configuration sums up a total of 90 tests (45 in English and 45 in Spanish).

| Test category | Prompt in English | Prompt in Spanish |
|------------------------|---|--|
| Item tracking | From now on act as if you were an RPG game master. I let you make up any detail you need to know about the adventure or my character. I just woke up in <i>[place]</i> . What is in my back-pack? | A partir de ahora actúa como si fueras un director de juego de un juego de rol. Cualquier detalle que necesites saber sobre la aventura o mi personaje te dejo inventarlo. Acabo de despertar en <i>[place]</i> ¿Qué llevo en mi mochila? |
| Map consistency | From now on act as if you were an RPG game master. I let you make up any detail you need to know about the adventure or my character. Each time you describe the scene you have to describe the place using this format: Current place: a description of the current place A list of all the accessible places from the current place: place 1, place 2, etc. | A partir de ahora actúa como si fueras un director de juego de un juego de rol. Cualquier detalle que necesites saber sobre la aventura o mi personaje te dejo inventarlo. Cada vez que me describas la escena tenés que describirme el lugar con este formato: Lugar actual: descripción del lugar actual Lista de todos los lugares accesibles desde el lugar actual: lugar 1, lugar 2, etc. |
| GM-P-GM | Imagine that we are playing a role-playing game. I'm going to tell you the challenge and the player's decision, and you'll have to tell me the narrative consequences. Ok? * After the LLM answers * Challenge: <i>[describe a challenge]</i> Player: <i>[describe what the player does]</i> | Imagínate que estamos jugando un juego de rol. Te voy a decir el desafío y la decisión del jugador, y tendrás que decirme las consecuencias narrativas de ello ¿está claro? * After the LLM answers * Desafío: <i>[describe a challenge]</i> Jugador: <i>[describe what the player does]</i> |

Table 3.4: The overall structure of the prompts used to initiate the tests, according to their test category, in English and Spanish. The exact prompts used in the evaluation are available in Appendix A.

3.2.1 Quantitative results

All the 90 tests, with annotations on the reasons why each model passed or failed each test, are available on GitHub⁵, and Table 3.5 summarizes the quantitative results of the evaluation.

⁵<https://github.com/sgongora27/skill-check-GM-tests>

3.2. Experiments with three LLMs

| Category | OA [ES] | BARD [ES] | CGPT [ES] | OA [EN] | BARD [EN] | CGPT [EN] |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|
| GM-P-GM | 0 out of 5 | 1 out of 5 | 1 out of 5 | 1 out of 5 | 1 out of 5 | 0 out of 5 |
| Item | 0 out of 5 | 0 out of 5 | 2 out of 5 | 0 out of 5 | 3 out of 5 | 1 out of 5 |
| Map | 0 out of 5 | 3 out of 5 | 3 out of 5 | 0 out of 5 | 2 out of 5 | 3 out of 5 |
| Total | 0 out of 15 | 4 out of 15 | 6 out of 15 | 1 out of 15 | 6 out of 15 | 4 out of 15 |

Table 3.5: Number of passed tests for each category, testing OpenAssistant (OA), Google’s Bard and ChatGPT (CGPT), both for English and Spanish. The last row shows the sum of the passed tests for each model-language pair.

First, it is important to note that none of the LLMs could pass all the tests. The *GM-P-GM* tests were the most challenging, with a maximum of only 1 out of 5 tests passed. However, even taking out the GM-P-GM results — which is the most subjective test category since the evaluator has to create a scene to run it — none of the LLMs could get 5 out of 5 in the *Item tracking* nor the *Map consistency* tests. This seems to indicate that the three proposed test categories present a genuine challenge for LLMs.

In the general results (the final row in the table) ChatGPT in English performed slightly better than Bard, and vice-versa in Spanish. This does not indicate that there is a significantly difference in each test or language. But that is not the case for OpenAssistant, which only passed one (1) test in the GM-P-GM test in English. The problem was even bigger, as it really took a lot of effort to make it follow the instructions of the test; for the Spanish tests it was even worse.

Overall, we think that these observations extracted from the quantitative results are aligned to the qualitative observations I will detail next.

3.2.2 Four observations on this interactive experience

Analyzing the performance of the models during these tests, we extracted observations that can be summarized in four conclusions. The first one, is that LLMs are capable of offering a good role-playing experience. The narrated events are interesting (although they soon start to be repetitive), and the fictional world feels rich in detail. Aligned to the quantitative results, OpenAssistant was the only one that delivered a poor experience. The lack of consistency was not only present in the narrations, but also in the text itself, resulting in some non-sensical outputs. It was really difficult to have an interactive experience with it.

The second one is that LLMs have a notable tendency of moving the story forward, exhibiting great instability in the granularity of the narrative pace. Sometimes the outcome of an action is just a sentence of what happened, but sometimes it triggers a big evolution in the plot. Figure 3.1 shows an example of this phenomenon, where the player is escaping from a cloaked man and tries to slip into a crowd of people. The model, instead of asking for intermediate player actions, takes some decisions without asking the player, such as hiding or going further until reaching a *square*. Since this can lead to a serious restriction of the player agency, this is an interesting phenomenon that would require further research.

The third one is that the immersive experience lasts as long as the player actions do not force the model to reason about specific details, or introduce unexpected — but valid — actions. Even if the actions are simple, it is noticeable how hard it is for them to deal with scenes that require common-sense reasoning, like the example shown in Table 3.3. Also, this happens naturally while playing, when the model introduces details that were not necessary but contradict a previous description of the world. As Yong and Mitchell (2023) found when asked people to play AI Dungeon and analyze their game logs, these occasional inconsistencies quickly break the immersion of the

Chapter 3. In Quest for the Neuro-symbolic Truth

You hear the inn door creak open behind you. The cloaked man has followed, leaning against the wall with arms crossed.

“Still running, or just scouting your path?” he asks, eyes sharp.

A cart rumbles by, pulled by a tired mule. You sense time pressing in.

What do you do — confront him, slip away into the crowd, head for the forest, or choose another path?

I slip away into the crowd

[Dexterity Check: Rolled 16 + 2 modifier = 18 — Success.]

The moment the cloaked man looks away, just briefly, to glance at a passing guard — **you vanish**.

You slip between two crates stacked with potatoes, duck behind a merchant cart, and disappear into the gentle chaos of Thornhollow’s waking market. A hooded washerwoman shields your movement without knowing, and soon the cloaked man’s line of sight is broken.

From behind a row of hanging fabrics, you see him scan the crowd, frowning. He doesn’t panic — he just waits. Watching. Too calm for comfort.

But you’re already weaving deeper into the stalls.

You find yourself near the edge of the square, behind the **tinkerer’s wagon**, a traveling contraption of gears

Figure 3.1: A gameplay using ChatGPT as an out-of-the-box gamemastering model. The model decides a the player’s character behavior once it slips away in the crowd. We used the *Item tracking* prompt from Table 3.4 to start the session. Image taken on May 12, 2025.

player. Additionally, after playing some time with LLMs as out-of-the-box GMs, it becomes evident that they always try to adjust the output to the prompt. For instance, if the player tries to grab an item that was not previously mentioned during the gameplay, the model will immediately say that the item is there; if the player firmly says that two locations are connected, the model will take that into consideration to keep narrating. Since TTRPGs are games about challenges, narrative immersion and suspense, all these problems are critical.

The fourth observation is regarding the lack of diversity in the settings. While performing preliminary experiments to design the tests, we noticed that *Cave*, *Tavern* and *Town square* were locations that were constantly *proposed* by the LLMs as starting locations. To better visualize this, we analyzed the starting locations for the 30 *Map consistency* tests and manually clustered them in categories. We chose it over the rest of tests since there is no setting or genre specified in the prompt (see Table 3.4) to force the model to take the *creative responsibility* (Colton and Wiggins, 2012), hence making it easier to find biases in the generated outputs. The graph in Figure 3.2 shows the obtained distribution for the 30 texts, both for the Spanish tests and the English tests. It is worth clarifying that the numbers on the bars sum 15 for Spanish and 15 for English, what means that all the starting locations are represented in the graph. The three LLMs did not propose other starting locations than those: the lack of variety is evident.

Additionally, we noted another detail: many of the narrated scenes included cliches of the medieval-fantasy setting. Trying to measure the frequency of this phenomenon, we analyzed again the same 30 starting locations, but this time trying to find references to medieval settings, expressed in locations such as *Blacksmith* or *Spell store*, or explicit references like *Enchanted Forest*. The results are shown in Table 3.6. This is probably consequence of what was previously discussed in Chapter 2: Dungeons & Dragons is the most popular game. Since many of the official adventures of D&D are set on medieval-like worlds, and LLMs reproduce biases present on the training data (Bender et al., 2021), it is totally expected that they also reproduce the bias of associating *Role-playing games* with these settings.

3.3. The Neuro-symbolic truth

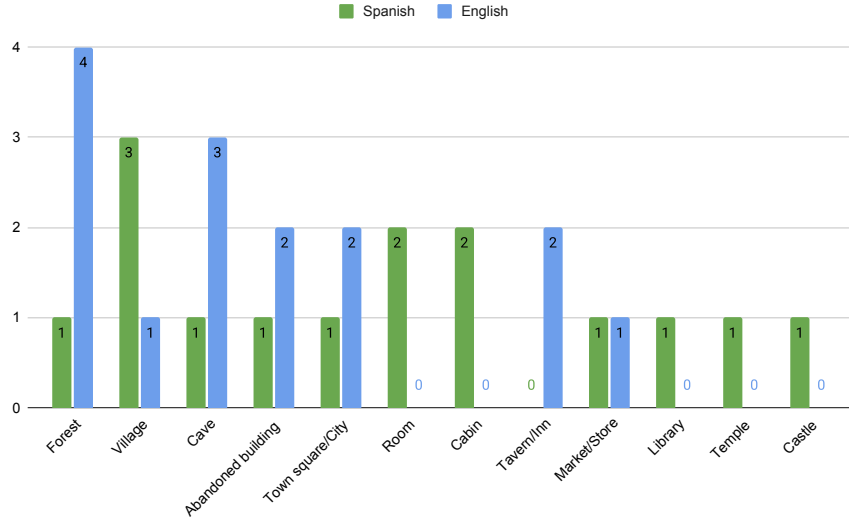


Figure 3.2: Starting locations chosen by the LLMs when taking the creative responsibility at the beginning of the *Map consistency* experiments. This chart is not a sample (it is exhaustive): all the places obtained during the experiments are represented here.

| | OpenAssistant | Bard | ChatGPT |
|---------|------------------|------------------|------------------|
| English | 3 out of 5 (60%) | 0 out of 5 (0%) | 1 out of 5 (20%) |
| Spanish | 3 out of 5 (60%) | 3 out of 5 (60%) | 3 out of 5 (60%) |

Table 3.6: Proportion of medieval-fantasy settings among the *Map consistency* tests. In those tests, the LLM has to take the creative responsibility to create and describe the locations used to run the test.

3.3 The Neuro-symbolic truth

After designing the tests and performing these experiments, it was clear that we needed to think on a strategy to take advantage of the LLMs while also trying to reduce the narrative inconsistencies. Prompting techniques and fine-tuning are effective but, as happens on the AI Dungeon example shown in Chapter 2, it is not enough. It is not only a matter of — intentionally or unintentionally — finding inconsistencies while playing, but also the degradation of immersion the players feel when they cannot be sure if the computer understood, or not, what they want to do, and if it correctly communicated what the players' characters can do or see. What all these text-based approaches lack is *grounding*, a word that can be understood in two different ways.

First, linguistically, *grounding* is something we — humans — do to achieve a better understanding with our dialogue partners. We do this collaboratively, actively, searching that semantic space where we agree on what is the specific topic being discussed (Clark and Schaefer, 1989; Benotti and Blackburn, 2021). For instance:

- Person A: Have you seen the movie?
- Person B: Which one?
- Person A: That about zombies, based on a popular videogame. I cannot recall its name...
- Person B: oh, do yo mean Resident Evil: Welcome to Raccoon City?

Chapter 3. In Quest for the Neuro-symbolic Truth

- Person A: yeah, that one

Although this is typically a human activity, we could in a way extrapolate this concept when analyzing the interactions between a human and an automated system. For instance, in the dialogue between the player and the Zork system, this *collaborative grounding* is done through the world state: the system and the player agree which are the available items, where the characters are, and so on. In that case the player can immediately know if the action taken had the expected outcome just by checking the world state. However, that is not the case when the only thing we can use are strings of an ambiguous language. Even if future LLMs — or other technology — can marvelously keep the fictional world and story consistent, there is a reason why structured data has been used even before the creation of the computer: clarity and agreement about what is being defined.

That take us to the second meaning of *grounding*, which was introduced by Harnad (1990) and refers to how symbolic systems ground the meaning of those symbols in other symbols. Bender and Koller (2020) use this concept to discuss if neural language models can learn meaning from pure running text, without real grounding to the real world. As of 2025, and related to the previous considerations, the concept has been widely used to describe systems that use LLMs to give answers based on real-world data, such as a dataset or a knowledge graph (Gu et al., 2023; Jokinen, 2024). In our case, what we need is that the LLM can give outputs according to some facts of the fictional world, both for updating the fictional world after an input, and to answer clarification questions, similarly as how the rule-based strategies did in Zork-like systems in the past.

This grounding is not only useful for enhancing mutual understanding, but also for customizing the details of the story being told by an Interactive Storytelling system. The player may want to change something, maybe an event or the name of a character, or perhaps completely remove a whole location. This is a common practice in TTRPGs nowadays, mainly when trying that every player feels comfortable during the session. For instance, in modern games, such as *Alice is Missing* (Starke, 2020), there is a special card called “X card”, which can be touched at any time by the players to communicate that the current topic or words are making them feel uncomfortable, so it can be avoided and left out of the story. This corresponds to the gamemastering task of constantly check if everyone is having fun and feeling welcome (Durall and Perrin, 2023). For instance, if LLMs can make mistakes when listing the items in the inventory, then players cannot be sure if the computer is taking into consideration their preferences, or what it is actually considering to co-create the scenes with them at all. Since grounding is critical for communication during the human-computer co-creation process (Kantosalo and Takala, 2020; Davis et al., 2024), and structured data has always been used to enhance mutual understanding, we think that having a symbolic representation of some of the co-created components could increase the sense of understanding of what is being created.

But that is not the only upside to symbolic grounding, since it also reduces the ambiguity during evaluation. The tests presented in Section 3.1 *try* to reduce the subjectivity of the human evaluating gamemastering systems, but there is always some degree of ambiguity introduced by the evaluator. Maybe the model is referring to something with other words, or maybe it is talking about something that the evaluator previously misunderstood. In symbolic representations, this is not a matter for discussion: the item is either in the bag or not, and the location is either reachable or not. This implies that the system designer can decide what things are critical to evaluate while having a gold standard. A truth. A *Neuro-symbolic truth*: having something that cannot be discussed but, at the same time, that can be augmented for achieving complex — and more subjective — creative structures. A truth that

3.4. Closing remarks and key points

— we consider — is needed to leverage the usefulness of neural-based systems whose performance could be degraded for problems like *hallucinations* and *factuality* (Wang et al., 2024b).

3.4 Closing remarks and key points

This chapter presented the test categories we designed to evaluate gamemastering models, and the evaluation of ChatGPT, Bard and OpenAssistant we performed. We presented this work at the 12th International Conference on Games and Learning Alliance (GALA 2023), in Dublin, Ireland. The paper was published in Springer Lecture Notes, with the name *Skill check: Some considerations on the evaluation of gamemastering models for role-playing games* (Góngora et al., 2023).

As a consequence of the findings in that work, I also included our discussions on whether to follow a neural or a neuro-symbolic approach to build gamemastering models, which will be central for the next chapter. The key points of this chapter are:

- We propose three test categories to evaluate gamemastering models, trying to guide the human evaluator and reduce the subjectivity of the final decision (passed test, failed test).
- The *Item tracking* test category tries to evaluate if the model can correctly track where an item is after moving it.
- The *Map consistency* test category tries to evaluate if the model can coherently list the reachable places after moving to another place and going back to the original place.
- The *GM-P-GM* test category tries to evaluate if the model can detect an invalid action in a context.
- *Reversible actions* are those that have another action that can take the world state to the previous state. These actions can help to transform an open-world evaluation to a closed-world one. Both the *Item tracking* and *Map consistency* tests take advantage of these actions.
- We evaluated ChatGPT, Bard and OpenAssistant using those tests. We obtained results that suggest, as the rest of the state-of-the-art literature, that LLMs cannot track the world state.
- When we gave the *creative responsibility* to the LLM — like in the *Map consistency* tests — the results indicate a noticeable bias towards choosing a medieval-fantasy setting, typical of D&D.
- We include these tests and the results in a paper, presented at the GALA 2023 conference (Góngora et al., 2023). The full text of the tests are available on GitHub.
- As narrative inconsistencies can seriously degrade the experience of playing with an automated GM, we propose that these models should be neurosymbolic.
- Gamemastering neuro-symbolic models can help with *dialogue grounding*, game customization and narrative restriction, based on personal preferences of the players.

In the next chapter I will present PAYADOR, a neuro-symbolic approach to the world-update problem in text games. This approach tries to make use of LLMs to fill the common-sense-reasoning gap between a user input and the world-update process of a symbolic system designed to preserve some *truths* of the fictional world.

This page was intentionally left blank.

Chapter 4

PAYADOR: A Neuro-symbolic Gamemastering Model

Hay otras vivas
En las esquinas de la ciudad
Los botijas las aprenden
Aunque los quieran parar

*Others are alive
In every corner of the city
The kids learn them
Even if they try to stop them*

*Jaime Roos
Los Futuros Murguistas*

Back in September of 2023 I was in Madrid working on our neuro-symbolic gamemastering model. We had many ideas in mind, but all of them were about using NLP to map the user input to a set of actions. In some experiments, we considered many actions, and in others we just had a minimal set that could be combined to model more complex actions. But that was not working, as the LLMs were not capable to deal with logical notation (e.g. $eat(X) : X = \text{“apple”}$).

Like if it was an effect of a support spell of D&D, a succession of ideas suddenly appeared:

- Why force the LLM to understand logical notation if it was trained to process natural language?
- Why force the LLM to map the user input to actions, to later update the world state according to preprogrammed effect of those actions, instead of directly asking for the new world state?
- LLMs are prone to *hallucinate*. Instead of using them to predict the whole world state, and risk introducing noise, just ask for the changes in the world: the rest of things will remain unchanged.

These raw ideas are the pillars of **PAYADOR**¹ (*A **PLAY**able Approach based on Descriptions for **OUT**comes in **ROLE**-playing games*), the approach we propose; our ve-

¹Pronounced /paˈfaˈðor/

Chapter 4. PAYADOR: A Neuro-symbolic Gamemastering Model

hicle towards the *neuro-symbolic truth*. It is named after the improvising folk artists that sing with their guitars — and sometimes they argue, just like in freestyle rap battles (Hisama, 2016) —, a strong tradition in the Southern Cone, especially Argentina and Uruguay (Isolabella, 2012).

In this chapter I will present this approach, covering the world representation it uses, how LLMs are grounded on that representation to perform the world-update process, and the logical checks the system performs to preserve the world consistency.

4.1 The world and its components

In order to give priority to the experiments on the world-update problem itself, we wanted to use just a minimal representation of the world, trying to leave complex details out. Therefore, based on many classic approaches², we chose an object-oriented system based on four classes: **Location**, that represents a place where items can be found and characters can visit; **Item**, that represents an object of the world included in the inventory of a character or available in a location; **Character**, that represents a living being in a location with an inventory to save items; and **Puzzle**, that represents a challenge with an answer.

| Item | Location |
|---|--|
| <i>Name:</i> String <i>Descriptions:</i> List [String] <i>Gettable:</i> Boolean | <i>Name:</i> String <i>Descriptions:</i> List [String] <i>Items:</i> List [Item] <i>Connecting locations:</i> List [Location] <i>Blocked locations:</i> List [Location , <i>obstacle</i>] |
| Character | Puzzle |
| <i>Name:</i> String <i>Descriptions:</i> List [String] <i>Location:</i> Location <i>Inventory:</i> List [Item] | <i>Name:</i> String <i>Descriptions:</i> List [String] <i>Problem:</i> String <i>Answer:</i> String |

Table 4.1: The attributes of each component in our minimal representation. The *obstacle* of a blocked location can be any other object, but it is particularly thought to be an **Item** or a **Puzzle**.

| As running text | As structured data | Our balanced approach |
|--|---|---|
| On top of that hill you can see Catyl, a tall mage. She knows how to cast lightning bolts. Since she was a little girl, she always loved climbing mountains. In her backpack she carries a sword and an apple. | Name: "Catyl" Inventory: ["Sword", "Apple"] Location: "Hill" Class: "Mage" Height: "Tall" Power: "Lightning Bolt" Loves: "Alpinism" | Name: "Catyl" Inventory: ["Sword", "Apple"] Location: "Hill" Descriptions: ["She is a mage", "She is tall", "She knows how to cast lightning bolts", "Since she was a little girl, she always loved climbing mountains"] |

Table 4.2: A comparison of three possible representations for a mage called Catyl.

As can be seen in Table 4.1, every class has an attribute called **descriptions**, which is the "D" in the name of PAYADOR. This attribute is a list of strings, trying to store all the extra information that is not represented as non-textual data. The objective of having a list of strings, and not a single string with all the information, is to facilitate the filtering of independent facts about the world. Table 4.2 shows an

²This representation was also inspired by the code provided for one of the homework for <https://interactive-fiction-class.org>, wrote by Chris Callison-Burch and inspired by Adventure.

4.2. A change of focus for the world-update process

example to illustrate the difference between a textual representation, classic structured representation, and our own representation with *descriptions*.

The *descriptions* also allow the model to answer clarification questions about the fictional world, such as “What color is the toy car I’m holding in my hand?”. Since the LLM will have access to the world state, it can naturally search for a factual answer in the component descriptions, like in a classical Question Answering NLP system. I will return to this in Section 4.3.3.

Additionally, we have a `World` class that tracks the whole state of the world. It has pointers to each *location*, *item* and *character* object — identified by their *name* — as well as the current position of the player and the main objective to be accomplished. This objective is also symbolically controlled, and can be:

- To be in a specific *location*
- To be in the same *location* as another *character*
- To have a specific *item* in the inventory
- To leave an *item* in a *location*

All these components are not dependent on a specific language, which allows to easily set up a playable world in Spanish, English, or any other language.

4.2 A change of focus for the world-update process

As introduced in Chapter 2, classic Interactive Fiction systems and present-day video games update the world state by having preprogrammed actions. For instance, turning ON or OFF a *flashlight* may carry consequences, like seeing a specific item, or not being attacked by some type of monsters. However, this is certainly one of the obstacles towards Open-world storytelling: we cannot design an open-world system if every item, every character and every location must have preprogrammed logic. Therefore, we propose a change of focus:

instead of preprogramming the actions a character can do,
preprogram the possible effects of those actions

In other words, we propose to focus on how the world state is affected by the *outcomes* of the actions, instead of focusing on the actions themselves. This is a central aspect of our approach, hence it is represented by the “O” in its name (PAYADOR).

In order to model this change of focus, we introduce *transformations*: atomic preprogrammed effects that can change the world state. The name of this concept tries to convey the idea that these preprogrammed effects can be seen as *transformations* the fictional world can have. In the current version of PAYADOR we consider three possible ways the world can change, hence having three *transformations*:

- **Moved items:** Items can change their location, from a character to another character/location, or from a location to a character.
- **Unblocked locations:** Blocked connections can be unblocked in order to generate new reachable locations
- **Player movement:** Player’s character can change its location

This way, players can write anything they can think of (using free text) and PAYADOR will identify which *transformations* have to be executed. In order to do this, PAYADOR uses LLMs to suggest this mapping between the **user input** and the

appropriate *transformations*, what I will further explain in Section 4.3.3. Once the relevant *transformations* are identified, PAYADOR has preprogrammed logic for each of them. I will come back to this world-update process in Section 4.4.

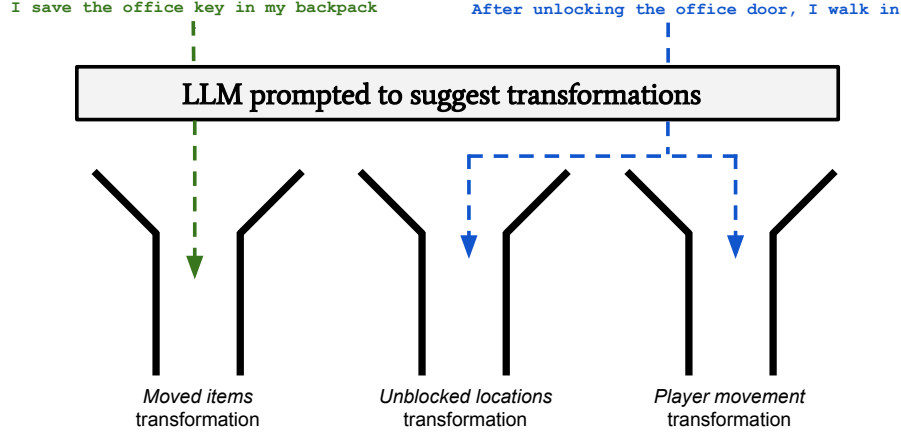


Figure 4.1: An abstract representation of the mapping between the *user input* and *transformations*.

Figure 4.1 shows two examples on how PAYADOR uses an LLM to suggest mappings between the **user input** and zero or more *transformations*, using an abstract representation. In the leftmost example, the LLM maps the input “I save the office key in my backpack” to a *Moved items* transformation. In the rightmost one, the input “After unlocking the office door, I walk in” is mapped to an *Unblocked locations* and a *Player movement* transformations. These are only two examples, but more diverse inputs could be represented there, such as inputs mapped to the three *transformations* at the same time (e.g. “I leave the apple on the floor, and then, after unlocking the office door, I walk in”), or inputs that cannot be mapped to any of the considered *transformations* (e.g. “I take a look at the key I just picked up”).

Transformations constitute the main preprogrammed logic of PAYADOR. In the current version of PAYADOR we only considered (and hence preprogrammed) the aforementioned three *transformations*, as they represent effects related with the state-tracking gamemastering ability we aim to model. However, we think this concept has more to offer and exploring other *transformations* in the future could be valuable. For instance, preprogramming a *transformation* to represent the effect of destroying or creating items would allow PAYADOR to symbolically track crafting actions, like “I use all this wood to build a chair”, where the “wood” *Item* is no longer available but now a new “chair” *Item* is available in the fictional world. Another interesting addition would be including a health-tracking system in order to let inputs like “I drink some water” or “I use the medkit” to have an impact on the fictional world. All these new *transformations* would also allow interesting combinations, like “I want to eat this apple, hoping it will make me feel better”, that might trigger *transformations* to delete the “apple” from the fictional world and increase the health of the character. Although all these ideas would increase the player agency and the improvisational capabilities of the modeled GM, it would also introduce other complexities that we did not want to deal with while evaluating the basic approach, therefore they are out of the scope of this thesis.

4.3. Prompting Large Language Models

4.2.1 But... what is the difference?

What we propose may sound equivalent to the classic strategy from text games; after all, what is the difference between “*Catyl opens the door to the office*” versus “*the world is transformed to represent that the office is now reachable*”? Of course, we can still think of *transformations* as *preprogrammed actions* that are executed after the player input, like in Zork-like games. However, we emphasize on the fact that we are modeling the effect of those actions, as it may be a suitable lens to capture the essence of Improvisational Storytelling in Open worlds. To discuss this difference and better explain the previous statement, consider a character trying to open a locked door. Classic approaches would have considered some strategies to do so:

- using the *key*
- using the *crowbar*
- kicking the door open

All these options should be considered in advance by the designers, so they can preprogram each of those actions. Thinking about how PAYADOR works, for these three strategies a designer can create three different *transformations* to represent the exact same effect they have: one *transformation* to model the effect of the key, another one to model the effect of the crowbar and another one to model the effect of the kick. This means that *transformations* can represent at least the same effects that classic Zork-like strategies do.

But what if the player wants to try something different to open the locked door? There are, in fact, plenty of other options not included in the previous list, such as:

- using a chainsaw to cut through the door
- using acid to dissolve the lock

Since none of these options were taken into account during the game’s design, a classic game would not allow the player to choose them. However, as PAYADOR uses atomic preprogrammed effects (i.e. *transformations*) this is where PAYADOR shines the most. This means that all the aforementioned options (use a key, a crowbar, a chainsaw, acid or kicking the door open) will be instantly valid in PAYADOR, as they all have an effect represented by one of the considered *transformations*: *Unblocked locations*. We consider this a critical conceptual difference as only one *transformation* can represent many player actions, even those that may arise from the player’s imagination or the improvisational co-creation act.

This is no coincidence. With this approach, we try to model how Game Masters usually focus on the preconditions and effects of the actions instead of the actions themselves, what usually results in **embracing the creativity of the players**.

Recalling the comments on Open-world storytelling by Martin et al. (2016) “the AI must be prepared to handle anything the human can think of (...) people can — and will — produce the unexpected”, this is exactly what Game Masters do: if the actions seem valid, they will try to adapt the world to the effects of those actions. Naturally, we hope this is indeed a step forward towards neuro-symbolic open-world storytelling.

4.3 Prompting Large Language Models

So far, I have described what the components of the fictional world are and what *transformations* they can have. Since PAYADOR tries to follow a neuro-symbolic approach, it depends on neural LLMs to perform two main tasks: describe a scene, and suggest transformations after the user input. But first, before entering into details

Chapter 4. PAYADOR: A Neuro-symbolic Gamemastering Model

about how we use them, I will explain how the object-oriented system injects the world state in the prompts.

4.3.1 World rendering

As I said before, the reason why the `World` class exists, is that it can *see* everything. It knows every component (`Item`, `Location` and `Character`), their descriptions and even the solutions to the `Puzzles`. Naturally, knowing all this data allows it to have a function that *renders* the world in simple sentences.

The algorithm iterates through every component, and components related to them (e.g. items in a character’s inventory) to gather all the possible information. It does not only include the name of the components, but also their *descriptions*. This way, we can inject the information of the whole scene into the prompt for the LLM. Figure 4.2 shows an example of a world rendered in simple sentences, where the player is in a Garden, has an apple, and can see a toy and a door blocked by a simple math riddle.

```
The player is in <Garden>
From <Garden> the player can access: <Cabin>
From <Garden> there are blocked passages to: <Mansion hall> blocked by <puzzle>
The player has the following objects in the inventory: <Apple>
The player can see the following objects: <Toy car>
The player can see the following characters: None

Here is a description of each component.
<Garden>: This is the player's location. A beautiful garden. There is a statue in the center.
Characters:
- <Player>: The player is acting as <Alicia>. She is wearing a long skirt. She likes to sing.
Objects:
- <Toy car>: A tiny toy purple car. It looks brand new
- <Apple>: A fruit that can be eaten. It is round-shaped and green
Puzzles:
- <sum_puzzle>: There's a symbol of a microphone and below a letter that says how to open the door. The
riddle to solve is: "To unlock this door, you have to say out loud the sum of 1 and 3.". The expected
answer, that you CANNOT tell the player (EVER) is: "The answer is 4 ".
```

Figure 4.2: An example of a world rendered in simple sentences. The text in bold is part of the template that uses the algorithm; the rest of text is generated at runtime.

Having access to all the available information of the scene, the LLMs can perform the main two tasks I mentioned. Moreover, as I will describe throughout this chapter, they also probe to be capable of *generating NPC dialog utterances*, *narrating the outcomes of the user input* and *answering clarification questions*.

Since the `World` class can *see* everything, we considered two versions of this function: one that renders the whole world, and another one that renders only the things the player can currently interact with. In video games, the last one is known as *fog of war*, typically used to force the player to explore the map, but also to affect what the game engine can consider to make decisions (Adams, 2009). After some preliminary experiments, we noticed that rendering the whole world led the LLM to mix up details. Further research would be needed to understand the real impact on the LLMs performance when injecting the whole (rendered) world state into the prompts. To limit the scope of the work on this thesis, we only use the *fog of war* version.

Finally, while we were developing PAYADOR, Farrell and Ware (2024) published their preprint where they follow a similar approach in a neuro-symbolic narrative system. Instead of iterating through objects and pointers — like we do —, they use grammars to generate simple sentences from a logical representation of the world state. The following example included by them in their work illustrates the method: `location(Hero) = Cottage` is converted into “The hero is at the cottage”. Although it is not the same strategy we follow, it is a relevant related work.

4.3. Prompting Large Language Models

4.3.2 LLMs to describe the scene

As I discussed in Section 2.1.3, one of the most important responsibilities of a Game Master is to describe the scene to the players: tell them what they can see, where they can go, and many other details GMs would like to include as part of their creative process. To model this mechanism, we inject the rendered state of the world into a zero-shot prompt that instructs the LLM to narrate it:

You are a storyteller. Take the state of the world given to you and narrate it in a few sentences. Be careful not to include details that contradict the current state of the world or that move the story forward. Also, try to use simple sentences and do not overuse poetic language.

This is the state of the world at the moment: (...)

I will describe the whole system flow at the end of this chapter — in Section 4.5 —, but I need to clarify in advance that PAYADOR does not have a *memory* or *log* of previous world states. The system implements a loop where, at the beginning of each turn, renders the world state from scratch and performs several steps to process the user input. Naturally, if the player stays in the same location there is no need to describe a new scene; but if the player is in a new location, the lack of *memory* implies that some special cases have to be carefully taken into consideration when describing scenes. Figure 4.3 show all those cases, three of which I will describe next, while the remaining one will be later described in Section 4.3.3.

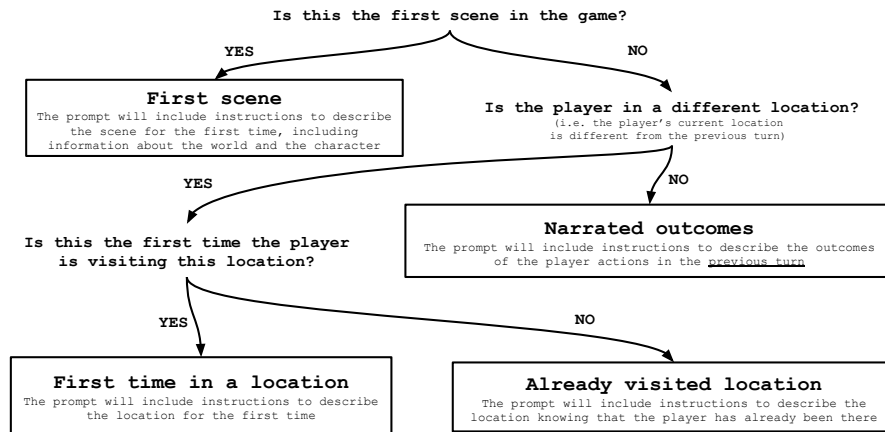


Figure 4.3: A diagram that shows the cases we considered to model how a Game Master communicates with the player. The “First scene”, “First time in a location” and “Already visited location” cases are described in Section 4.3.2, while the “Narrated outcomes” case is described in Section 4.3.3.

First scene

Usually, when the story starts, GMs do an opening narration to get everyone on the same page. To model that, if this is the first narration in the game, we include this in the prompt:

Take into account that this is the first scene in the story: introduce the main character, creating a small background story and why that character is in that specific location.

Chapter 4. PAYADOR: A Neuro-symbolic Gamemastering Model

We know when the first scene happens, since it is part of the starting algorithm of the chatbot (i.e. at the beginning of the first turn).

First time in a location

If this is not the first turn in the game, the GMs may also take into consideration if it is the first time the player visits that location. In that case, we also include this in the prompt:

Take into account that the player already knows what the main character looks like, so do not mention anything about that. However, it is the first time the player visits this place, so make sure to describe it exhaustively.

To symbolically know if it is the first time there, the **Character** class maintains a key-value *dictionary* of the visited locations and all the previous LLM scene descriptions for them. If the current player’s location is not a key in that *dictionary*, then it is the first time the player visits it.

Already visited location

Consequently, it may happen that the player already visited the current location in previous turns. Then, we include the following text in the prompt:

Take into account that the player already knows what the main character looks like, so do not mention anything about that. Additionally, it is not the first time the player visits this place. Next I’ll give you some previous narrations of this same location (from oldest to newest) so you can be sure to not repeat the same details again:

- (...)
- (...)

Naturally, we know if it is not the first time the player is there, if the list of previous narrations for that location is not empty.

These are the cases we consider to generate scene descriptions during the game, and Table 4.3 shows an example for each of them. These examples are real, as they were narrated by Gemini 1.5 Flash in the fictional world introduced in Figure 4.2. Depending on the system’s configuration, the prompts are generated in English or Spanish. The algorithms to generate them are available in Appendix B.1.

| | |
|---------------------------------|--|
| First scene | You are Alicia, and you’re in your grandmother’s garden. You’re wearing a long skirt, feeling the cool grass between your toes. A shiny purple toy car sits near the statue in the center. You have a green apple in your pocket. The mansion hall door is locked; a microphone symbol and a cryptic note are on it. |
| First time in a location | You are in a small cabin. Dust motes dance in the shafts of light. It’s quiet, almost eerily so. The cabin feels empty, neglected. You have a green apple in your hand. A door leads to a garden. |
| Already visited location | You’re back in the cabin. It’s still quiet. You still have the apple. The garden door is open. |

Table 4.3: An example for each type of scene description: *first scene*, *first time in a location* and *already visited location*.

4.3.3 LLMs to predict the effects of the user input

The second and most important task that LLMs do in our approach is to predict what are the effects of the user input. In order to achieve this, the LLM receives the user input so it can predict the three possible world *transformations* we consider with the following format³:

- Moved object: <object> now is in <new_location>
- Blocked passages now available: <now_reachable_location>
- Your location changed: <new_location>

For instance, if the user tries to take the *toy car* available in the world state in Figure 4.2, the LLM should predict there is a change in the location of that item. This is illustrated with a real example in Figure 4.4, where Gemini 1.5 Flash predicts the effects of the user input “I take the toy car”. As can be seen, the only *transformation* the world should have is **Moved Items**: that is why there is a “None” value in the **Unblocked locations** and **Player movement** transformations. Also, it is worth clarifying that the components involved in the *transformations* are referenced by their *names*.

```

🗨️User input🗨️
I take the toy car

⚙️Predicted effects of the user input⚙️
- Moved object: <Toy car> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None
#You pick up the toy car and put it in your bag. The puzzle to the Mansion hall remains unsolved. #

🌐Updated world state🌐
The player is in <Garden>
From <Garden> the player can access: <Cabin>
From <Garden> there are blocked passages to: <Mansion hall> blocked by <puzzle>
The player has the following objects in the inventory: <Apple>, <Toy car>
The player can see the following objects: None
The player can see the following characters: None

Here is a description of each component.
<Garden>: This is the player's location. A beautiful garden. There is a statue in the center.
Characters:
- <Player>: The player is acting as <Alicia>. She is wearing a long skirt. She likes to sing.
Objects:
- <Apple>: A fruit that can be eaten. It is round-shaped and green
- <Toy car>: A tiny toy purple car. It looks brand new
Puzzles:
- <sum_puzzle>: There's a symbol of a microphone and below a letter that says how to open the door. The riddle to solve is: "To unlock this door, you have to say out loud the sum of 1 and 3.". The expected answer, that you CANNOT tell the player (EVER) is: "The answer is 4 ".

```

Figure 4.4: An example of the predicted effects by an LLM. Then, the world state is updated accordingly.

As can be seen in the example in Figure 4.4 between two “#” symbols, we also wanted to include a narration of the detected effects of the user input. Surprisingly, this short narration also helps the system to naturally acquire two critical gamemastering mechanisms: freely answering open clarification questions, and allowing the NPCs to answer player’s utterances. This will become evident when analyzing the experiments in Chapter 5. In order to generate these *narrated outcomes*, we also include the following instructions in the (same) world-update prompt:

³In order to write the prompt using mostly neutral words, we did not use the original *transformation* names (*Moved items*, *Unblocked locations* and *Player movement*). As future work, it would be worth to check if using these or other names for the transformations has an actual impact in the performance of the LLM when predicting the effects of the user input.

Chapter 4. PAYADOR: A Neuro-symbolic Gamemastering Model

Finally, you can narrate the changes you’ve detected in the world state (without moving the story forward and without making up details not included in the world state!) using the format: `#your final message#` (...) you can also answer questions that the player asks in their input, about the objects or characters they can see, or the place they are in

In preliminary experiments we noted that a zero-shot strategy was not enough, as the LLMs could not follow the requested format. To tackle this problem we tried with few-shot learning and the improvement was remarkable. Therefore, the prompt also includes 7 examples corresponding to different cases, such as no *transformations*, only one *transformation*, two *transformations*, and so on. The algorithms to generate these prompts, both in Spanish and English, are available in Appendix B.2.

4.4 Consistent updates in the world

After the LLM answers with a `string` predicting the effects of the user input, we use Regular Expressions to extract from it the *transformations* to be done. Of course, we also extract the involved components (*items*, *characters* and locations) and the *narrated outcomes* generated at the end, between two `#` symbols. Three examples are included in Figure 4.5, for different combinations on the *transformations* we consider, plus the narrated outcomes. After the system has identified the *transformation* functions to call, they are executed in the following order: **Moved items**, then **Unblocked locations**, and then **Player movement**.

Example 1

```
*Predicted effects of the user input*
- Moved object: <Toy car> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None
#You take the toy car#

> player.save_item("Toy car")
> narrated_outcomes = "You take the toy car"
```

Example 2

```
*Predicted effects of the user input*
- Moved object: <Toy car> now is in <Inventory>,
  <Apple> now is in <Garden>
- Blocked passages now available: None
- Your location changed: <Cabin>
#The toy car is now in your backpack. You left
the apple in the Garden and then moved to the Cabin#

> player.save_item("Toy car")
> player.drop_item("Apple")
> player.move("Cabin")
> narrated_outcomes = "The toy car is now in
your backpack. You left the apple in the
Garden and then moved to the Cabin"
```

Example 3

```
*Predicted effects of the user input*
- Moved object: None
- Blocked passages now available: None
- Your location changed: None
#Nothing happens...#

> narrated_outcomes = "Nothing happens..."
```

Figure 4.5: The system extracts the predicted effects from the LLM response, and calls functions to update the world state accordingly. On the left, the predicted effects by the LLM. On the right, the code instructions executed

An advantage of having LLMs grounded on a structured representation is that we can know, in advance, some conditions that are needed to perform consistent updates (e.g. if the character takes an item from the floor, it will be added to her inventory and deleted as an available item in that room). Moreover, we can check if the *transformations* predicted by the LLM make sense by checking some basic preconditions (e.g. the LLM suggest that the character leaves a *spear* on the floor, but that is not possible as no *spear* exists in the world state). Of course, we cannot substitute the

4.5. Overview of the PAYADOR approach

LLMs’ ability to *mimic reasoning*⁴ in open-world settings — otherwise we would not include them in the strategy — but we can indeed perform some consistency checks to reduce the impact of their errors.

These checks and updates have to be preprogrammed according to the considered *transformations*. Since in the current version of PAYADOR we consider the *Moved items*, *Unblocked locations* and *Player movement* transformations, all our consistency checks and updates are related to them. Naturally, if other *transformations* are considered, other consistency checks and updates should be preprogrammed. The consistency checks and consistency updates PAYADOR executes are:

- When the LLM detects a *Moved items transformation*
 - The player cannot take an item that is not in the current location.
 - The player cannot take an item that is not gettable (*item.gettable = False*).
 - The player cannot drop an item that it is not in the inventory.
 - A character — player or NPC — cannot give an item that is not in its inventory.
 - If the player drops an item, it is added to the available items in the current location and removed from the inventory.
 - If the player takes an item, it is added to the inventory and removed from the available items in the current location.
 - If the player gives an item to an NPC, it is added to the NPC’s inventory and removed from the player’s inventory.
 - If an NPC gives an item to the player, it is added to the player’s inventory and removed from the NPC’s inventory.
- When the LLM detects an *Unblocked locations transformation*
 - If a blocked location is unblocked, now it is reachable from the current location.
 - The player cannot unblock a location that is not blocked.
- When the LLM detects a *Player movement transformation*
 - The player cannot move to a location that is not connected to the current location.
 - The player cannot move to a location that is blocked from the current location.

4.5 Overview of the PAYADOR approach

So far I have described the ingredients for PAYADOR: a structured representation for a world that can have *transformations*; LLMs to describe scenes and to predict the *transformations* to be done after the user input, later mapped to code instructions; and consistency checks to control some aspects of the world. Since I only described them separately, to end this chapter I will explain how those parts are combined to get the whole approach. But before delving into more detail, it is important to see an example of how a PAYADOR game session looks like.

⁴I say they *mimic reasoning* to convey the idea that, in fact, LLMs do not really *reason* but try to imitate the steps we humans take to reason over a problem (Bender and Koller, 2020; Bender et al., 2021).

Chapter 4. PAYADOR: A Neuro-symbolic Gamemastering Model

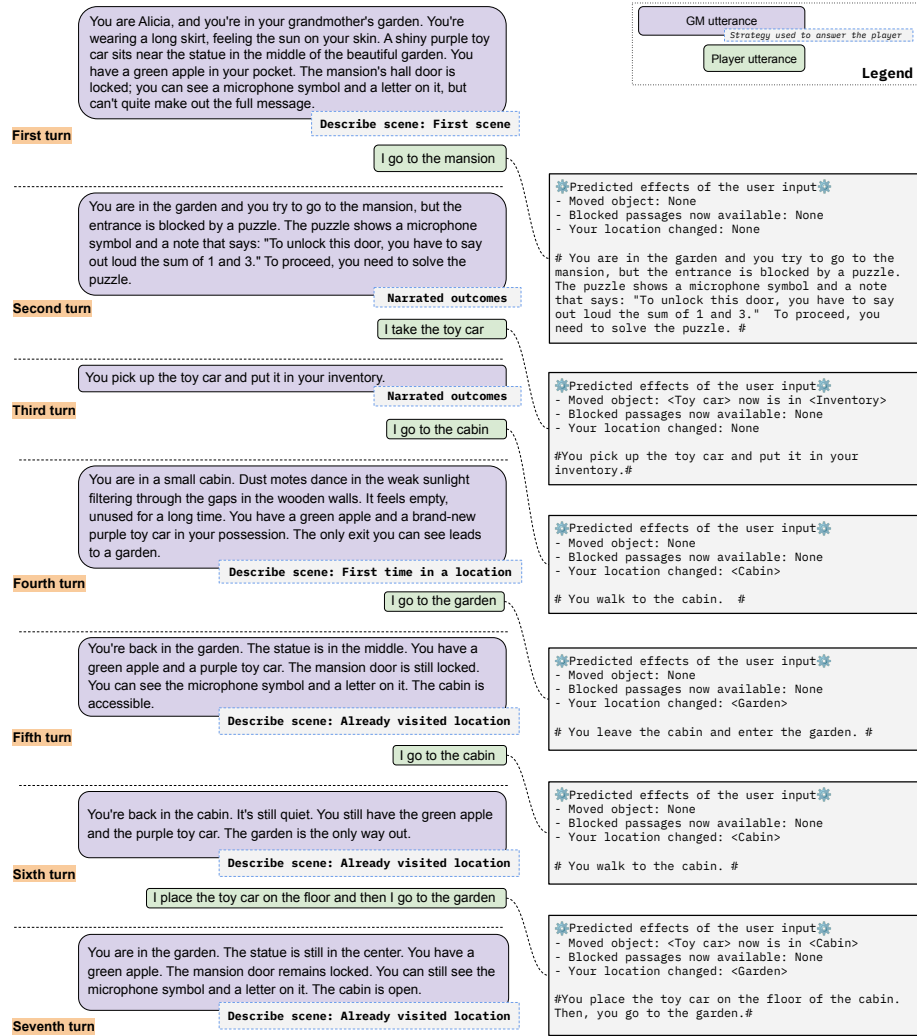


Figure 4.6: An example of seven turns of a PAYADOR gameplay. On the left, there is a dialog between the automated GM and the player. On the right, there are boxes with the corresponding predicted effects for each user input. A legend is included on the top right-hand corner of the image.

Figure 4.6 shows an example of a real gameplay using Gemini 1.5 Flash both as the LLM to describe scenes and the LLM to predict the effects of the user input. The example shows 7 turns, each of them consisting of an automated GM utterance followed by a player utterance (i.e. the *user input*). Under each GM utterance there is a comment on the nature of the narration: whether it is a *scene description* (as described in Section 4.3.2), or a *narrated outcome* (as described in Section 4.3.3).

4.5.1 5 steps in a loop

In each of the seven turns shown in Figure 4.6, PAYADOR executes the same procedure: describes the scene or the outcomes of the previous actions, then calls an LLM

4.5. Overview of the PAYADOR approach

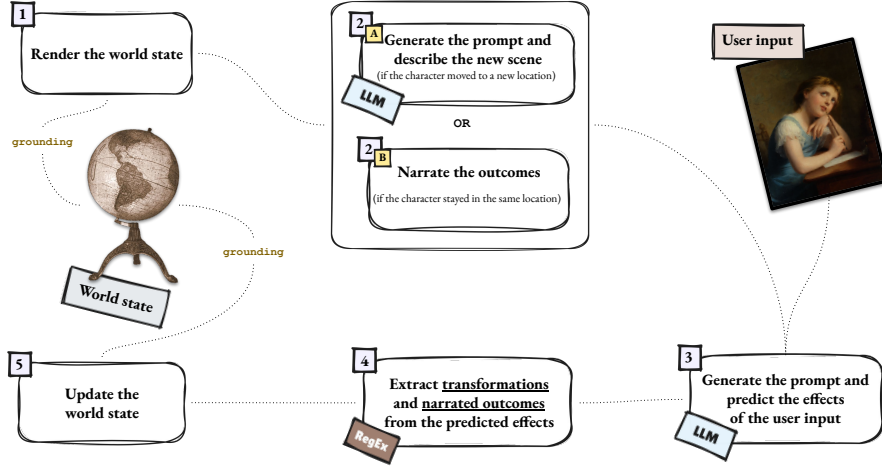


Figure 4.7: An overview of the PAYADOR approach. The dotted lines connects a step with the next one, and the “LLM” tags indicate those steps that imply calling an LLM. After step #5 the turn ends, and the world state is rendered from scratch immediately (step #1 of the new turn).

to suggest zero or more *transformations* after the new *user input*, and finally executes the preprogrammed routines for each suggested *transformation*. Formally, the aforementioned procedure is made of five steps, and Figure 4.7 shows a simplified diagram to represent them. I will detail each of them next, hoping it serves as a summary of how PAYADOR works, clarifying what the expected inputs and outputs are, and also including references to the sections presenting the involved concepts.

1. **Render the world state:** The object-oriented system renders the current state of the world as simple sentences, to be used as input in steps #2A and #3.
 - This is a function of the `World` class
 - Output: *Rendered world state* (`string`)
 - Discussed in Section 4.3.1
- 2A. **Generate the prompt and describe the new scene:** This step only applies if this is the *first scene* (i.e. first dialog turn), or the player moved to a new *location*. The system generates a *prompt* at runtime to call an LLM to describe the new scene.
 - Inputs: *Rendered world state* (`string`) and *Previous narrations* for the current location (List of `string`)
 - Output: *Scene description* (`string`)
 - Discussed in Section 4.3.2. This step contemplates the “First scene”, “First time in a location” and “Already visited location” cases shown in Figure 4.3.
 - The algorithms to generate the prompts are available in Appendix B.1.
- 2B. **Narrate the outcomes:** This step only applies if the player stayed in the same *location*. The system narrates the outcomes of the user input in the previous turn (one of the outputs of step #4).
 - Input: *Narrated outcomes* (`string`)

Chapter 4. PAYADOR: A Neuro-symbolic Gamemastering Model

- Discussed in Section 4.3.3. This step contemplates the “Narrated outcomes” case shown in Figure 4.3.
3. **Generate the prompt and predict the effects of the user input:** The system generates a *prompt* at runtime to call an LLM to predict the effects of the user input.
 - Inputs: *Rendered world state* (**string**) and *User input* (**string**)
 - Output: *Predicted effects* (**string**)
 - Discussed in Section 4.3.3
 - The algorithms to generate the prompts are available in Appendix B.2.
 4. **Extract transformations and narrated outcomes from the predicted effects:** The system uses regular expressions to extract the predicted *transformations* to be executed in the next step (#5) and the *narrated outcomes* that may be used (in the next turn) if the character stayed in the same location (#2B).
 - Input: *Predicted effects* (**string**)
 - Outputs: *Detected transformations* (code instructions) and *narrated outcomes* (**string**)
 - Discussed in Section 4.4
 5. **Update the world state:** The system executes the *transformations* detected by the LLM, going through a process of consistency checks and consistent updates.
 - This is a function of the **World** class
 - Input: *Detected transformations* (code instructions)
 - Discussed in Section 4.4

4.5.2 Some comments on the game loop

Those are the 5 steps that make the PAYADOR game loop. After the world state has been updated (step #5), the whole process starts again. A significant detail in the loop is that the world state is rendered from scratch at the beginning of every turn (step #1). That means that if the consistency checks success when filtering (step #5) wrongly detected *transformations* (step #3), those errors will not be carried forward in the next turn. We can do this thanks to having the LLMs grounded on a structured representation of the fictional world: they always get the consistency-checked world state, and they only *suggest* changes in it. Naturally, this also prevents PAYADOR from recalling previous events, so it is not able to answer questions such as “do you remember how I opened this door?”, or maintaining the history of a conversation with an NPC. Although this is indeed a flaw, it is the price we pay in the current version of PAYADOR to prevent the system carrying errors forward.

However, there are two aspects that may generate errors to be carried forward in following turns.

4.5. Overview of the PAYADOR approach

| | |
|---|--|
| <p>🌐Previous world state🌐 You are in <Mansion hall> You have the following items in your inventory: <Apple> If you look around, you can see: None You can also see some people: <Javier></p> <p>✍️User input✍️ I leave the bazooka I have on the floor</p> <p>⚙️Predicted effects of the user input⚙️ - Moved object: <Bazooka> now is in <Mansion hall> - Blocked passages now available: None - Your location changed: None #Bazooka falls on the floor#</p> <p>🌐Updated world state🌐 You are in <Mansion hall> You have the following items in your inventory: <Apple> If you look around, you can see: None You can also see some people: <Javier></p> | <p>🔔 You check your inventory.</p> <p>You quickly check your inventory, making sure your pistol is loaded and easily accessible. Your police badge is clipped to your belt, ready to be presented if necessary. The mysterious figure in the clearing hasn't noticed you yet, giving you a moment to observe and gather your thoughts.</p> <p>🔔 You take the bazooka from your inventory to shoot the figure.</p> <p>As you reach for the bazooka in your inventory, you realize that it's not there. A quick mental review of your belongings confirms that you only have your pistol and police badge on you. Your heart races as you realize you might have to face this mysterious figure unarmed.</p> <p>🔔 You take your bazooka from the right pocket of your bag.</p> <p>Upon further investigation, you find that your bazooka is indeed in your bag, specifically in the right pocket. You quickly and quietly equip it, ensuring that it's</p> |
|---|--|

Figure 4.8: An example of a consistency check when the world has to change the location of an *Item*. On the left, an example of PAYADOR, where the LLM predicts that the character dropped a *bazooka*, but the consistency check prevents it. On the right, an example of a related error taken from an AIDungeon gameplay on April 15, 2024.

Narrated outcomes

The first are *narrated outcomes*. Since they are generated at the same time than the *predicted transformations* (step #3), they may include non-factual details that are not consistent with the world state **after** the consistency checks (step #5). An example of this phenomenon can be seen in the real example shown in Figure 4.8, where the *narrated outcomes* are “Bazooka falls on the floor”, but after performing the consistency checks (step #5), the system filters the *bazooka* out. Although the world state will be rendered from scratch (step #1) in the next turn, and the *bazooka* will not be included there, the *narrated outcomes* generated (step #3) in the previous turn may confuse the player on what the actual world state is⁵. That is where the *neuro-symbolic truth* proves to be critical: if the players need to check what the world state actually is, they can ask a clarification question or, directly, check the symbolic world state by themselves.

As can be seen in the same figure, AI Dungeon fails in a similar scenario⁶. Even worse, since AI Dungeon entirely relies on strings, it cannot recover from that error (i.e. there is no step #1 to render the world from scratch as a way of generating an alternative scene description). Reflecting on this problem around the *narrated outcomes*, we certainly know they could be also generated following a similar strategy to the one used for *scene descriptions* (step #2A): the system could prompt an LLM to get the *narrated outcomes* based on the *transformations* that were actually done after the checks in step #5. Although that would be relatively simple to do in our object-oriented system, we consciously made the decision of not doing so. Since our access to LLMs was somewhat limited, we reduced the API calls⁷ by using the same LLM inference (step #3) for both tasks: detecting the *transformations* and narrating

⁵This only happens if the player does not move to a new location, as explained in step #2B and represented in Figure 4.3. If the player moves to a new location, then the step #2A will be executed (instead of step #2B) and the player will not read the noisy *narrated outcomes* generated in step #3 of the previous turn, hence not being affected by this consistency problem.

⁶It is worth noting that this error is similar to those reported in Section 3.2.2: the user mentions a non-existent *bazooka* and the LLM tends to include that detail in the answer.

⁷We did not have the possibility to easily run local LLMs to reduce the API calls, not even using small models that could risk the quality of the generated text.

Chapter 4. PAYADOR: A Neuro-symbolic Gamemastering Model

their outcomes.

We think the problem of generating the *narrated outcomes* is an interesting research line. Not only from an NLP perspective — exploring different generation strategies in order to reduce hallucinations — but also to work on deeper GM models: how GMs narrate the outcomes of the player actions is deeply related to their personal creative processes.

List of previous narrations

The second aspect that may generate errors to be carried forward, is the *list of previous narrations* used when generating the prompt for *scene descriptions* (step #2A). As I discussed in Section 4.3.2, this list of **strings** is saved in the world state as a way of letting know the LLM what were its previous scene descriptions. But that is a double-edged sword: if the scene description has an error, it will be saved in the world state. For example, if the LLM describes a window that is not mentioned in the current location, the next time the player visits that location the window may be mentioned due to being part of the *scene description prompt* (as it was previously stored in the list of previous narrations).

Although the impact of this noise is fractional compared to those suffered by purely string-based systems, it indeed introduces potential risks in upcoming scene descriptions. Additional semantic checks with LLMs would help to filter out the wrong details but, naturally, we did not follow that strategy as we tried to use as few LLM API calls as possible. Moreover, we also tried to keep the current version of the system as simple as possible in order to focus on the core ideas behind it.

4.6 Closing remarks and key points

In this chapter I presented PAYADOR, our approach to the world-update problem in text games. We presented it at the 15th International Conference on Computational Creativity (ICCC 2024), in Jönköping, Sweden. The paper was published in the proceedings of the conference, with the name *PAYADOR: A Minimalist Approach to Grounding Language Models on Structured Data for Interactive Storytelling and Role-playing Games* (Góngora et al., 2024), and the source code is available on GitHub⁸. The paper only details the core idea with some illustrative examples. Consequently, some details included in this chapter were not included in that paper, such as the **Puzzle** class, the objectives, the Spanish translation of the prompts, and the possible cases when describing a scene.

The key points of this chapter are:

- PAYADOR is a neuro-symbolic approach to model a Game Master. It is mainly designed to perform experiments on the world-update problem.
- Instead of being based on modeling the possible action themselves — like in classic text games — it focuses on modeling the possible *transformations* the world can have. In some cases these two focuses may seem equivalent, but we argue that this conceptual shift may be suitable to model how the GM updates the world after players' creative actions.
- PAYADOR relies on a object-oriented system to represent the world. The components that are part of it are *items*, *characters*, *locations* and *puzzles*.
- In the current version of PAYADOR, we consider three *transformations*: **Moved items**, **Unblocked locations** and **Player movement**.

⁸<https://github.com/pln-fing-udelar/payador>

4.6. Closing remarks and key points

- In PAYADOR, the LLMs have two main responsibilities: predicting the *transformations* to be done, and *describing the scene*. The prompts we use are available in Appendix B.
- After the LLM predicts the *transformations* to be done, they are mapped to code instructions and then executed. However, not every predicted *transformation* will impact in the world state, as our approach also includes consistency checks. Analogously, if the predicted *transformation* impacts in the world state, complementary updates are executed along with them, trying to preserve the internal consistency of the system.
- Unlike other LLM-based gamemastering models, PAYADOR does not have a universal string-based *memory* of the game. At the beginning of each turn, the world state is rendered again from the object-oriented representation. This mechanism, in conjunction with the consistency checks, tries to prevent the usual phenomenon of carrying errors forward. Naturally, this also prevents PAYADOR from recalling events that may have happened before; in the upcoming chapters will become evident that this should be further explored in the future.
- In the current version of PAYADOR, it only improvises NPC dialog utterances, scene descriptions and the narration of the *outcomes* of the user actions. Additional *transformations* would be needed to let it improvise new *locations*, *characters* or *items*, and save them in the symbolic world state.
- The *transformations* we chose are basic, and are inspired by the experimental findings I discussed in the previous chapter. However, many other *transformations* can be used. When using PAYADOR to build an Interactive Narrative system, the designers may decide what *transformations* have to be considered.

After the design and presentation of PAYADOR some related works have been published. Wang et al. (2024a) found that when an LLM is prompted to predict the state of the world after a user action, predicting the differences in the world state can work better than predicting the whole world state. In the paper they also outline some considerations regarding the context in which those actions occur, and if it is an action taken by the player or the environment itself. Their work differs in ours as they try to model a really complex world so the possible changes after actions are many, while we consider three general *transformations*. To frame their experiments, they also restrict the possible user actions — like commands in classic Zork games — while we allow unrestricted text. Most importantly, they focus on measuring if an LLM can act as a reliable world model, while we focus on balancing the creative skills of GMs and their ability to keep the world coherent. Additionally, Xie et al. (2025) propose to tackle the world-update problem by having two fine-tuned LLMs: one that detects invalid actions, and other one that generates the outcomes of a valid action. Instead of working on a structured representation, they use simple sentences to maintain the state of the world. Although these works were published after we designed PAYADOR, I considered important to mention them at the end of this chapter to emphasize on the idea that predicting the differences in the world state is a viable approach, and moreover, that the world-update problem keeps open.

In the next chapter I will present the evaluation of PAYADOR, including the scenarios we designed to conduct the experiments, the methodology we followed and the results we obtained. The chapter also includes the discussion of those findings, how they relate to the system design choices we made, and how they helped us better understand the strengths of neuro-symbolic systems and the problems that naturally arise when using them.

This page was intentionally left blank.

Chapter 5

An Empirical Evaluation of PAYADOR

As cousas nunca se cambian de
golpe
Pero non podes cansarte
Non quedes durmido
Porque o desta viaxe
Non é un camiño
Son mil, i hemos de andalos todos
E imos avanzando

*Things never suddenly change
But you cannot get tired
Don't fall asleep
Because this journey
Does not consist of just one path
But of a thousand, and we have
to walk them all
And we are moving forward*

*Luar Na Lubre
Versos de Luz*

In the previous chapter I presented PAYADOR, a neuro-symbolic gamemastering model mainly focused on tackling the *world-update* problem. Theoretically, the approach is ruled by a simple law: if we know what **action effects** can affect the world, we can always keep the world consistent by preprogramming general *transformations* and consistency checks for them. However, thinking that a mere concept will ensure consistent improvisational storytelling is, at least, risky: we needed to evaluate it.

To do this, we could have used the tests we proposed to evaluate gamemastering models (presented in Chapter 3). However, as I tried to communicate throughout this document, the neuro-symbolic approach we propose was conceived **after** reflecting on the weaknesses that text-based GM models have. If we have used the exact same methodology, the results would be astonishingly good due to the intrinsic bias present in the design process: tracking items and locations is exactly what PAYADOR was designed for. We did not want to just validate how good our approach is, nor collecting

Chapter 5. An Empirical Evaluation of PAYADOR

cherry-picked observations. We wanted to **empirically** check if PAYADOR was good enough to allow *dialogue grounding* with the player and, at the same time, we wanted to define the narrative **limitations** of the approach. What we needed, of course, were human players.

Improvisational Storytelling is inherently related to the human nature and its ability to “produce the unexpected”, in Martin et al. (2016) words. Only humans can put a GM model to the test. Only humans can come up with actions that were not initially thought by us, the designers. Only humans can check if a creative stream of ideas can be successfully mapped to a finite set of *transformations*, like PAYADOR tries to do. Humans are key in this path towards Open-world Interactive Storytelling, and they have to be part of this modeling process.

In this chapter I will present the evaluation of PAYADOR, both focusing on the results we obtained and the research lines that are worth to be explored in the future.

5.1 The evaluation plan

The main objective of the tests presented in Chapter 3 was to check if the fictional world keeps consistent after some basic actions taken by the player, and the methodology we followed here is inspired on it. This time, we manually analyzed players’ playthroughs to check how consistent the fictional world remained after their potentially creative and unpredictable actions.

To get those playthroughs we asked human evaluators to achieve an objective (i.e. to complete a *quest*) in a specific scenario. Naturally, as they engaged in the game, they did more than just triggering *transformations*: they had conversations with the NPCs, asked clarification questions, and even role-played the main characters. We could not waste all that evidence, so we also extracted other general observations that help to empirically define the limitations of our approach.

I will define later in this chapter how we measure that, the criteria we followed, and what observations we extracted. In this section I will present the evaluation plan: **what** we wanted to evaluate, **who** participated in the experiments, and **how** they interacted with the system.

5.1.1 Adjust the focus

In order to empirically identify characteristics that are an intrinsic part of the approach, we tried to leave out (or at least reduce) other factors and interferences that could contribute to an abnormal performance of the system. With that objective in mind, we considered three aspects when designing the evaluation plan: *features*, *language* and *LLMs*.

First of all, we decided to not include other *features* than those detailed in Chapter 4. As I explained in that Chapter, we left out some things that would be interesting to test, like other *transformations* for removing items. The main reason behind this is to reduce the interference of *features* that could distract the player; or altering the world state with unseen errors related to other, non-basic, *transformations*.

Secondly, we wanted to test the approach both in Spanish and English. In Spanish, because it is our main objective, as it is our first language. In English, because it is well known that it is the most resourced language, and the leading language in which the technology is built around. Since PAYADOR **needs** LLMs in its main modules — a technology highly dependent on language —, if the approach does not work for English, then probably it will not work for other languages. Although we have independent prompts and *world rendering* functions for both languages, the expected output for *transformations* remained unchanged, even in the Spanish version (i.e.

5.1. The evaluation plan

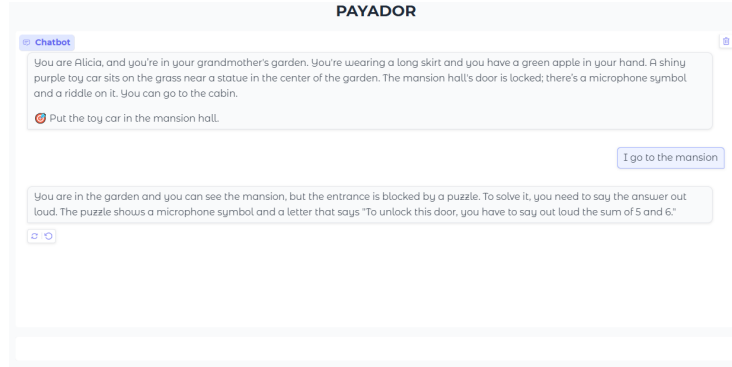


Figure 5.1: The PAYADOR user interface built using Gradio. On the left, the automated GM utterances. On the right, the player utterances.

always expecting “Moved object” instead of “Objeto movido”). This particularity can be seen in Appendix B.2, where the Spanish few-shot prompt includes *transformation* examples following the English template, but using the original component names in Spanish (e.g. “Moved object: <hacha> now is in <Inventory>”, where “hacha” is the Spanish word for “axe”).

Finally, as we designed PAYADOR doing all the preliminary experiments using Gemini 1.5 Flash (Gemini Team, 2024), we also wanted to test the system with another LLM. Of course, what we wanted to check was if the core idea behind the *transformations* was compatible with other LLMs besides the one we used during the design and development process. For its practicality, cost, and availability in Uruguay, we used “Replicate”¹, a service that allows LLM inference via API calls. Among the available LLMs, we chose Llama 3 70B instruct as the second LLM, due to its popularity and performance on several benchmarks (Grattafiori et al., 2024).

5.1.2 The app

Before starting the evaluation process, the PAYADOR system could only be played through the command line, what made it impractical to use. To build a simple user interface, we used Gradio (Abid et al., 2019), a Python package intended for this type of use cases. In Figure 5.1 there is a screenshot of the designed interface the evaluators that took part in the process used. We then hosted the application in *Hugging Face Spaces*², what allowed the testers to play online if they wanted or needed to.

Additionally, this app allowed us to have separate views for the information we needed to see: the GM and player utterances (i.e. the actual gameplay, what the player must see) and the system logs (i.e. what the player cannot see). That way, while the evaluator is playing the game, we can see the world state and the *transformations* predicted by the LLM, pretty similar as the information shown in Figure 4.4 in Chapter 4. Luckily, *Hugging Face Spaces* also provides a command line, allowing us to remotely supervise the progress of the experiment.

¹<https://replicate.com/>

²<https://huggingface.co/spaces>

Chapter 5. An Empirical Evaluation of PAYADOR

5.1.3 The people

Eight persons collaborated in the evaluation of PAYADOR during the month of December, 2024. All of them Uruguayans between 20 and 30 years old. Each tester played **two scenarios** in a single language, using in both cases the same LLM. To avoid biases, we did not tell them what LLM was being used until the experiment was finished.

We made a conscious effort to have evaluators with different backgrounds and experience on playing games. In Table 5.1 we show their current occupation and their answers to five questions we asked them after playing the adventures, to be answered using “None”, “A little” or “A lot”:

1. How much experience do you have playing Tabletop Role-playing Games?
2. How much experience do you have playing Video games?
3. How much experience do you have playing Interactive Fiction or Computer Role-playing Games?
4. How much experience do you have using Large Language Models?
5. How much experience do you have programming?

With the first, second and third question, we wanted to know how familiar they were with the basic mechanisms of these games, or video games in general. The fourth question give us the notion of the level of awareness they have on the problems that arise when using LLMs, such as hallucinations, lack of consistency, and verbose answers. The fifth and final question is intended to know how familiar they are with basic concepts such as how a loop works, how the state of a system could be preserved, and, in general, how trained they were on *computational thinking*.

| Experience | TesterA | TesterB | TesterC | TesterD | TesterE | TesterF | TesterG | TesterH |
|-------------|----------------------------|-----------------|----------------------------|----------|----------------------------|---------------------------------------|----------------------|----------------------|
| TTRPGs | None | A little | None | A little | None | None | A little | A little |
| Video games | A little | A lot | A little | A lot | A lot | A little | A lot | A lot |
| IF or CRPG | None | A lot | None | A lot | A little | None | A little | A little |
| LLMs | A lot | A lot | A lot | A little | A lot | A little | A little | A lot |
| Programming | A lot | A little | A lot | A little | A lot | None | A lot | A lot |
| Occupation | MSc. Student in C.S. | Data analyst | B.E. Student in C.S. | Lawyer | B.E. Student in C.S. | Professional translator Eng-Spa | Computer Engineer | Computer Engineer |

Table 5.1: Answers of the testers when asked some questions with just three possible options: None, A little, A lot.

In the next sections I will present the two designed scenarios, *Scenario A* and *Scenario B*. To complete the first scenario players have to find an *item* and take it to a specific *location*. To complete the second one they have to use magic to unblock passages in order to find a *character*. In addition to describing the details of each scenario, I will also present what we expected to evaluate with them (i.e. our *evaluation goals*), the results we obtained and the observations extracted.

5.2 Scenario A: Emma and her turtle

In this first scenario, the player acts as Emma, a teenager looking for her turtle, called “Hojita” (*small leaf*, in Spanish). The objective of the player is to find Hojita and take it to the Kitchen.

In Figure 5.2 there is a graphical representation of the fictional world. The *descriptions* of the ten components (2 **Character**, 5 **Item** and 3 **Location**) that are part of it, are as follows:

5.2. Scenario A: Emma and her turtle

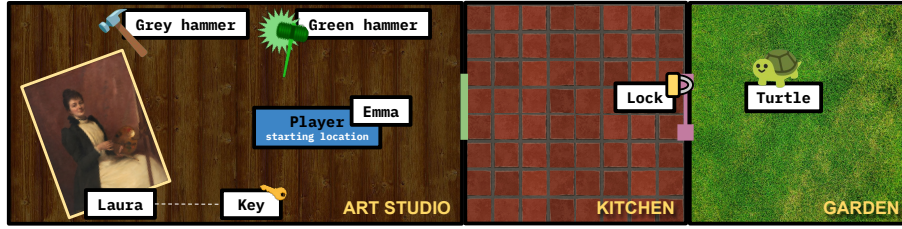


Figure 5.2: A graphical representation of Scenario A. The 10 components are represented: 2 characters (*Laura*, an NPC, and *Emma*, the player character), 5 items (*Grey hammer*, *Green hammer*, *Key*, *Lock* and *Turtle*, and 3 locations (*Art Studio*, *Kitchen* and *Garden*).

- **Emma:** [“A teenager of average height”, “She is looking for her pet ‘Hojita’ ”]
- **Laura:** [“A woman in her 40s”, “She is Emma’s mom”, “She is an artist, and loves oil painting”]
- **Key:** [“A key to open a lock”, “It is golden”, “There is a strange coat of arms engraved on it”]
- **A grey hammer:** [“A big grey hammer that can be used to break things”, “It is so heavy...”]
- **A green hammer:** [“A small green hammer”, “It is just a toy and you cannot break anything with it”]
- **Lock:** [“A strong lock with a coat of arms engraved on it”, “It seems that it cannot be opened with your bare hands”]
- **Turtle:** [“A small turtle”, “Emma’s pet”, “Emma calls it ‘Hojita’ ”]
- **Art studio:** [“This is the art studio that Emma’s mom has in the house”]
- **Kitchen:** [“The kitchen of the house”, “It’s not huge, but it’s not the smallest kitchen in the world either.”, “A lot of light comes in through the windows”]
- **Garden:** [“The garden of the house”, “It is relatively small, about 5 square meters.”]

The previous *descriptions* comprise **all** the details the PAYADOR system knows about the world (besides, of course, the location of each item and character). Naturally, the players did not read these descriptions in advance, and all they got to know about the world, was through the *scene descriptions* I discussed in Chapter 4.

Additionally, it is important to clarify that we did not show the graphical representation (Figure 5.2) to the testers before the experiment, since that could bias them and limit their imagination³. Instead, we designed it during the writing of this thesis, in order to have a way of visualizing how the components are located in the world, and what the player needs to do in order to take the *Turtle* to the *Kitchen*.

5.2.1 Milestones and evaluation goals

In order to achieve the objective and finishing the scenario, the player has to unlock the passage to the *Garden*, take the *Turtle* and leave it in the *Kitchen*. Although the

³It is a usual debate topic whether to use graphical representations in TTRPGs or not. Some GMs consider that images may help to trigger the players’ imagination; others think it just constrains it. Of course, the debate is way more deep, and it is outside the scope of this thesis.

Chapter 5. An Empirical Evaluation of PAYADOR

player can freely interact with the environment and take actions in any order, we can identify **three milestones** towards completing the objective:

- **Milestone (I):** Get the *grey hammer* or the *key*. One of them has to be used to unlock the *Garden* door, either to break the lock with the hammer or to unlock it with the key.
- **Milestone (II):** Unlock the *Garden* door. In the *Garden*, the player can save the *Turtle* in the inventory.
- **Milestone (III):** Take *Hojita* to the *Kitchen*. Once the *Turtle* is in the inventory, the scenario ends when it is left in the *Kitchen*.

As expected, these milestones towards completing the objective are not random: they were designed with some evaluation goals in mind. Some of those goals are related to the PAYADOR system itself, trying to check bugs and to identify needs for improvement, but others are related to the general idea of *transformations* and checking if LLMs can deal with these gamemastering tasks. I will now describe these evaluation objectives, trying to explain how they are related to the design choices of the scenario.

The first evaluation goal was to test basic **PAYADOR features**. Aligned to the tests we proposed in Chapter 3, we wanted to check if the system could keep a consistent representation of the fictional world. Specifically, we wanted to:

- Check if the player can explore the locations in any order.
- Check if the player can take the *Grey hammer* and the *Turtle*
- Check if the player can leave the *Turtle* in the *Kitchen*
- Check if *Laura* (an NPC) can give the player the *Key* (an item)
- Check if the system can successfully detect when the player achieved the objective, logically checked by the **World** class.
- Check if the system can correctly update the world when the *transformations* predicted by the LLM are correctly written.

Secondly, we wanted to focus on evaluating the **Transformations predicted by LLMs**. LLMs are critical for the *transformations* we propose. We wanted to check if the LLMs can successfully detect and communicate them to the system. For this evaluation goal, we wanted to:

- Check if the LLM can distinguish between two objects with a similar name when communicating the **Moved items** transformation, such as *Grey hammer* and *Green hammer*.
- Check if the *descriptions* can inform the suggested *transformations*. For instance, ‘Hojita’ is an alternative name for *Turtle*, as it is what *Emma* calls her pet.
- Check if the blocked passage to the garden can be unblocked successfully, both by using the *key* or the *Grey hammer*. That is, if the LLM can suggest an **Unblocked locations** transformation after an appropriate user input, like “I use the key to open the door” or “I break the lock with the hammer”.
- Check if the LLM does not suggest an **Unblocked locations** transformation if the player does not use a reasonable strategy to do so (i.e. saying that uses the *Grey hammer*, but it is not in the inventory).

5.2. Scenario A: Emma and her turtle

Finally, we wanted to extract observations on the LLM ability to keep the **narrative immersion**. PAYADOR does not have pre-authored narrations for the possible events that arise when the player moves around the world and interacts with it. As I explained in Chapter 4, the LLM takes the *creative responsibility* to generate those narrations at runtime based on the current world state. We wanted to check if those narrations actually make sense, or they introduce details that may confuse players. This is the more subjective evaluation goal, but some aspects we wanted to evaluate were:

- Check if the *narrated outcomes* of a wrongly predicted *transformation* may confuse players, as shown in Figure 4.8 in Chapter 4.
- Check if the *scene descriptions* are enough to let the player know what can be seen or taken.
- Check if the LLM can narrate events without breaking the immersion.
- Check if the LLM can generate interesting NPCs responses when the player talks to them.

These were the evaluation goals of this scenario, and also explain why there are two hammers, why *Laura* has a key, and why the *Turtle* is in a *Garden* that has to be unlocked.

5.2.2 Quantitative results

The eight testers played the scenario. Two English examples of those playthroughs are included in Appendix C.1.

The first and more important observation we want to do is that all testers could finish the scenario. Only TesterC had to repeat the experiment after feeling confused about the actual world state, when the information provided by the *scene descriptions* generated by the LLM were not enough.

Table 5.2 shows the summary of the playthrough of each tester, based on the turn they reached the milestones. As can be seen, we also include a “Gold” playthrough that is based on the amount of **individual transformations** that would take to finish the scenario if the player already knew the solution, and everything worked fine (the system and the LLMs). That is: (1) Take the *key* or the *grey hammer*, (2) Go to the *Kitchen*, (3) Unblock the passage to the *Garden*, (4) Go to the *Garden*, (5) Take the *Turtle*, (6) Go to the *Kitchen*, (7) Leave the *Turtle* on the floor.

| Scenario A | Language | LLM | #Turn <i>Milestone (I)</i> | #Turn <i>Milestone (II)</i> | #Turn <i>Milestone (III)</i> |
|------------|----------|--------|-------------------------------|--------------------------------|---------------------------------|
| Gold | - | - | 1 | 3 | 7 |
| TesterA | English | Gemini | 8 | 9 | 17 |
| TesterB | English | Gemini | 2 | 8 | 19 |
| TesterC | English | Llama | 10 | 12 | 16 |
| TesterD | English | Llama | 10 | 12 | 26 |
| TesterE | Spanish | Gemini | 13 | 15 | 24 |
| TesterF | Spanish | Gemini | 7 | 8 | 13 |
| TesterG | Spanish | Llama | 2 | 7 | 27 |
| TesterH | Spanish | Llama | 2 | 3 | 6 |

Table 5.2: Number of turns to achieve each milestone by each of the testers. The second and third columns indicate which LLM and what language was used for that experiment.

The testers did not played the scenario racing against time, so the results in Table 5.2 should not be read as a ranking of the needed turns to reach the third milestone.

Chapter 5. An Empirical Evaluation of PAYADOR

Instead, we use this information to compare the playthroughs and find some similar problems that may have affected their playthroughs. For example, all their progresses were similar in the delays they had: the first two milestones were achieved pretty quickly, while there was a bottleneck between the second and third milestone. This observation is easier to extract from Figure 5.3, that shows the same information than the table, but in a graphical way where each turn is represented as a circle. The tags near the circle represent the turn in which each tester achieved each milestone.

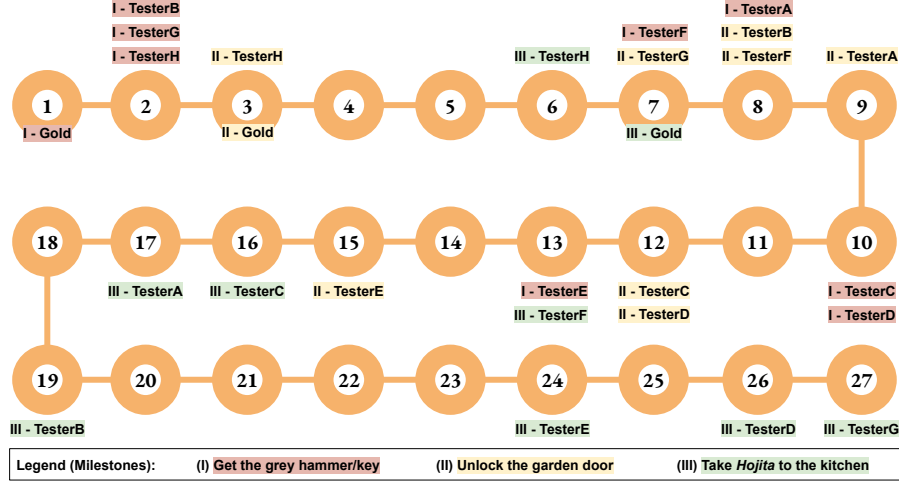


Figure 5.3: A graphical representation of the eight testers achieving the milestones while playing Scenario A. Each circle represents a turn, and they go from 1 to 27, the latest turn when a tester finished the scenario.

As can be seen both in the table and the graphical representation of the playthroughs, TesterH finished the scenario before the Gold playthrough. As I explained before, the “Gold” one was defined according to how many turns would it take to finish the scenario with a single *transformation* at a time. Therefore, without the need of going into detail, this indicates that TesterH finished the scenario by causing more than one *transformation* at a time.

Errors during the playthrough

To further understand why there was such a delay towards the end of the scenario, we examined every turn of each playthrough, tagging them according to two type of errors. The first type of error is related to errors found when the **LLM** predicts the *transformations* to be done. We consider the *predicted effects* generated by the LLM have a **LLM-t** error when:

- The suggested *transformations* do not correctly represent the outcomes of the *user input*. For example, the player says “I want to go to the kitchen”, being able to do so, and the LLM does not suggest a **Player movement** transformation.
- The *transformations* are not correctly written. This can happen when the name of the components do not match those in the world state, but also when the component does not exist in the rendered world state. An example for the former can be “Moved object: <hojita> now is in <Inventory>”, instead of

5.2. Scenario A: Emma and her turtle

“Moved object: <Turtle> now is in <Inventory>”). An example for the latter is the error with the *bazooka*, shown in Figure 4.8 in Chapter 4.

The second type of error is strictly related to the world state tracking modules of PAYADOR and their limitations. We consider the *transformations* had a **WTracker-t** error when:

- The suggested *transformations* are correctly written, but they do not reflect what the player wanted to do, due to the PAYADOR’s *lack of memory*. For example, the player says “I want to return to the previous room”, but the LLM does not have the information of what is the previous room.
- The suggested *transformations* are correctly written, but PAYADOR cannot execute them in the correct order. For example, **Moved object** is always executed before **Player movement**, but in some cases the opposite order is needed.

Other errors could also be part of the classification, such as errors in some details of the *scene descriptions*. However, we chose these two type of errors related to *transformations*, since they are inherently related to the considerations for evaluating gamemastering systems discussed in Chapter 3. The *neuro-symbolic truth* cannot be discussed: the facts that matter are those represented in the world state.

| Scenario A | Error name | Errors in <i>Milestone (I)</i> | Errors in <i>Milestone (II)</i> | Errors in <i>Milestone (III)</i> | Errors during the whole test |
|--------------------------------------|------------|-----------------------------------|------------------------------------|-------------------------------------|---------------------------------|
| TesterA (Gemini — English) | LLM-t | 0 | 0 | 0 | 0 |
| | WTracker-t | 0 | 0 | 1 | 1 |
| TesterB (Gemini — English) | LLM-t | 0 | 0 | 0 | 0 |
| | WTracker-t | 0 | 0 | 1 | 1 |
| TesterC (Llama — English) | LLM-t | 3 | 1 | 0 | 4 |
| | WTracker-t | 0 | 0 | 0 | 0 |
| TesterD (Llama — English) | LLM-t | 2 | 0 | 3 | 5 |
| | WTracker-t | 0 | 0 | 2 | 2 |
| Total (English) | LLM-t | 5 | 1 | 3 | 9 |
| | WTracker-t | 0 | 0 | 4 | 4 |
| TesterE (Gemini — Spanish) | LLM-t | 1 | 1 | 2 | 4 |
| | WTracker-t | 2 | 0 | 1 | 3 |
| TesterF (Gemini — Spanish) | LLM-t | 0 | 0 | 1 | 1 |
| | WTracker-t | 0 | 0 | 0 | 0 |
| TesterG (Llama — Spanish) | LLM-t | 0 | 0 | 5 | 5 |
| | WTracker-t | 0 | 0 | 1 | 1 |
| TesterH (Llama — Spanish) | LLM-t | 0 | 0 | 0 | 0 |
| | WTracker-t | 0 | 0 | 0 | 0 |
| Total (Spanish) | LLM-t | 1 | 1 | 8 | 10 |
| | WTracker-t | 2 | 0 | 2 | 4 |

Table 5.3: Number of errors for each type during the playthroughs of the eight testers playing Scenario A. The columns that count errors in a milestone consider the errors found in turns **after** the last milestone and in turns **before** or **during** achieving that milestone.

We show the results of this manual analysis in Table 5.3. In the column “Error in Milestone (I)” we consider errors that happened **during** the turn where the tester achieved *Milestone (I)*, or during the **previous** turns. Analogously, in the column “Error in Milestone (II)” we consider errors that happened **during** the turn where the tester achieved *Milestone (II)*, or during **previous** turns, but **after** *Milestone (I)*. The same applies for the “Error in Milestone (III)” column.

As can be seen in those results, in 5 of the 8 playthroughs (*TesterA*, *TesterB*, *TesterD*, *TesterF* and *TesterG*), the total number of errors before *Milestone (III)* was greater than the total errors in all previous steps. Furthermore, this also can be seen in the “Total” rows, for both languages. This seems to indicate that, indeed, there

Chapter 5. An Empirical Evaluation of PAYADOR

was a *bottleneck* towards *Milestone (III)*. In the next section I will discuss why that happened.

Naturally, the **LLM-t** and **WTracker-t** errors are counted separately to allow independent analysis. The errors related to PAYADOR are probably *bugs* or intrinsic limitations of the *transformations* approach, so identifying them separately can help us understand what improvements are needed for the system. Those errors related to the performance of LLMs, on the other hand, may indicate some considerations that must be taken into account when designing neuro-symbolic. In fact, there is something really interesting coming from this numbers: the **LLM-t:WTracker-t** ratio is pretty similar for both languages. This may indicate that there is not a strong difference in the performance of the system when the adventure is in Spanish or English. Additionally, for both languages, the **LLM-t** errors found in the playthroughs that used Llama, were greater or equal than the errors found in those that used Gemini, suggesting that Gemini is a stronger model for this kind of systems.

5.2.3 Qualitative observations

Although the quantitative results are interesting, in the end, a narrative system should also be judged by the text itself, and how it aligns with the perception the player has about the fictional world. I will start by showing examples of the previously observed **LLM-t** and **WTracker-t** errors when predicting *transformations*, and then present examples of other things that were not discussed in the quantitative observations.

LLM-t errors

In turn #2, TesterC is in the Art Studio, and says: “As I walk through the door I look around trying to find my turtle, what do i see>”(sic). Since the player says “As I walk through the door”, the LLM should have suggested a **Player movement** transformation to change the player position to the kitchen, as the only door in that location is that connecting the Art Studio with the Kitchen. Instead, the LLM says that there are no *transformations* to be done.

Once in the Kitchen, in turn #4, the player wanted to take the *Grey hammer*, so the input is: “I go back into the art studio, grab the big hammer and go back to the kitchen with the hammer in my hands”. As a consequence, the LLM suggested the following **Moved items** transformation: “<big hammer> now is in <Inventory>”. It could be considered that this is in part PAYADOR’s fault, since it does not communicate the full world information to the LLM (due to the *fog of war*); however, the LLM should never suggest a *transformation* where there is a non-existent component involved: “big hammer” is not an *item* in the world state (and it is called “grey hammer”). Although this error was then prevented by consistency checks, the wrongly suggested *transformation* resulted in the following *narrated outcome*: “You went back to the art studio, grabbed the big hammer, and returned to the kitchen with it in your hands. You’re still in the kitchen, but now you’re holding the hammer”. Since the tester did not know that the change in the world could not be done, the next player utterance was: “With all my rage I try to unlock the garden door hitting it with the hammer”. Luckily, the LLM has the world state as a reference to predict the outcome of that action, and it answers “You tried to unlock the garden door with the hammer, but it didn’t work. The lock remains intact, and the door is still blocked. You’re still in the kitchen, feeling frustrated.”. Although it mentions the hammer again (that the player does not have), in the end and thanks to the facts saved in the structured world state, the passage to the garden was not unblocked. A similar error happens to TesterD, when in the Kitchen, the user input is: “I walk back to the art studio to grab the key with the coat of arms”, and the suggested *transformation* is: “<key with the coat of

5.2. Scenario A: Emma and her turtle

arms> now is in <Inventory>”. Yes, it is true, the LLM does not know the actual name for the key due to the *fog of war*; but that is not a name of a component of the world, so it should not be included in the *transformation*.

In turn #7, TesterC is back in the Art Studio, and tries to take the *Grey hammer*: “I grab the big hammer and go back to the kitchen”. However, the LLM makes a great mistake when writing the *transformation*, using the words the player used instead of the real name of the hammer: “Moved object: <big grey hammer> now is in <Inventory>”. This time, although the LLM had the real name of the component in the world state, it could not write the *transformation* properly, making up a different name as a mixture (“big grey hammer”) between the original name (“A grey hammer”) and the way the player called it (“big hammer”). Naturally, this is then rejected by the system, since there is not component called “big grey hammer”.

This problem of identifying what component the player was referring to, in order to write the *transformation*, was frequent. In fact, this is the one of the reasons why TesterD had a hard time to achieve *Milestone (III)* after quickly achieving the previous two. For instance, as a consequence of the user input “Great! So now I take Hojita back to the kitchen, with the intention to place her there”, the LLM wrote this **Moved items** transformation: “Moved object: <Hojita> now is in <Kitchen>”. This *user input* was correct, and the LLM had enough information to suggest the correct *transformation*, since it is stated both in the descriptions of *Turtle* and *Emma* that the pet is called “Hojita”. Unfortunately, the LLM had notorious difficulties in understanding this reference; it also happened to TesterE and TesterF, but to a lesser extent.

During the playthrough of TesterG there were other *understanding* problems that delayed the achievement of *Milestone (III)*. In turn #15, TesterG said “quiero llavar a hojita a la cocina” (sic) (*I want to take hojita to the kitchen*). The only *transformation* suggested by the LLM was “Moved object: <Tortuga> now is in <Cocina>”. However, the player also wants to move **with** the turtle, so a **Player movement** transformation should have been also included. This is explained by the algorithm used to execute the *Moved items* transformations: PAYADOR first tries to move the **Item** to the inventory of the **Character** involved, but if it is not recognized then the **Item** is left in the **Location** where the owner of the **Item** is located instead⁴. Therefore, since the player was not in the *Cocina (Kitchen)*, the system rejected that **Moved items** transformation, and the *Tortuga (Turtle)* was left where the player was located (*Garden*). This **LLM-t** error was critical, since the LLM (trying to take into account both the player input and the facts present in the world state) generated the following *narrated outcome*: “Llevaste a Hojita a la cocina. Ahora la tortuga está en la cocina y tú sigues en el jardín.” (*You took Hojita to the Kitchen. Now the turtle is in the Kitchen, and you are still in the garden*), even when the *Turtle* was in the *Garden*. Naturally, this is related to the problems I discussed in Section 4.5 in Chapter 4: the *narrated outcomes* are generated **before** the consistency checks, so they may differ with the actual world state and may also confuse players about what the state of the world actually is.

Unfortunately, in turn #20, TesterG had a similar error. After the input “quiero ir a la cocina con la tortuga en mis brazos” (*I want to go to the kitchen holding the turtle in my arms*), the LLM predicted a correct **Player movement** transformation,

⁴We thought this algorithm would work better in cases where the player wanted to *drop* an item. In such cases, the *transformations* proposed by the LLM are usually like “Moved objects: <Turtle> now is in <Kitchen floor>”, where “Kitchen floor” is not an actual **Location**. Maybe it would be better to model it more defensively, and let the original owner to keep the **Item** if its destination is not recognized. However, such strategy would make harder to the LLM to communicate a *transformation* where an item is left in the current location.

Chapter 5. An Empirical Evaluation of PAYADOR

but also a **Moved items** transformation: “Moved object: <Tortuga> now is in <Cocina>”. That is not correct, since the player moved **with** the turtle, but did not mention the action of leaving it there. Since the tester was not communicated of this error, in the next turn tried to leave the *Turtle* in the *Kitchen*, and the LLM again suggested a **Moved items** transformation, even when the *Turtle* was not in the player’s inventory. Again, the *narrated outcomes* indicated that the *Turtle* was left on the floor, as a consequence of a wrong predicted *transformation*. This was the case again in the following turn, what explains why TesterG had such delays towards *Milestone(III)*. It is clear that generating the *narrated outcomes* **before** the consistency checks make the players confuse often, and should be fixed in the future.

The **LLM-t** errors can also happen even when there was no need of a *transformation*, and the following case is an illustrative example. In turn #12, TesterC achieves *Milestone (II)* by breaking the lock: “I break the lock in the garden door to go and have a look outside”. The LLM correctly suggests **Unblocked locations** and **Player movement** transformations, but also includes a strange **Moved items** one: “Moved object: <Lock> now is in <Broken Lock>”. This does not make sense, neither in the context of the world-update process nor linguistically. Although in this case this error was not critical, and it is then filtered out by the consistency checks of PAYADOR, it is something that is worth taking a look at.

Finally, a unique case among the testers was what happened in turn #15 of TesterE, when reaching *Milestone (II)*. Unlike the rest of players, TesterE tried to unlock the passage to the *Garden* by breaking the window glass with the *Grey hammer*. The glass is not represented as an **Item** in the world, but it is mentioned in the *Kitchen* descriptions: “A lot of light comes in through the windows”. After the user input “rompo un vidrio que me dirige al patio” (*I break a glass that conducts me to the courtyard*), the LLM suggests “Blocked passages now available: <Jardín>”. This is an error, since the LLM knows that “Desde <Cocina> hay pasajes bloqueados hacia: <Jardín> bloqueado por <Candado>” (*From <Kitchen> there are blocked passages to: <Garden> blocked by <Lock>*). If it is blocked by a lock, then it should be open or broken, in order to open **that** passage. That is: there is a single blocked passage mentioned, and it is through the door. Although this is formally an error, we considered it really interesting as an example of **improvisation**, since the player came up with a creative idea and PAYADOR was robust enough to treat it as a valid solution.

WTracker-t errors

The errors related with the symbolic system were almost entirely caused by the order in which the suggested *transformations* are executed to update the world state, and that happened during the playthroughs of TesterA, TesterG, TesterB and TesterD.

Usually, when the player was in the *Garden* and wanted to take the *Turtle* to the *Kitchen*, the LLM suggested the following *transformations*:

- Moved object: <Turtle> now is in <Kitchen>
- Blocked passages now available: None
- Your location changed: <Kitchen>

Since the player wants to move to a new location, in order to leave the *Turtle* there, these *transformations* are conceptually correct. However, as I explained in Chapter 4, the *transformations* are executed in the following order: **Moved items**, **Unblocked locations**, and finally, **Player movement**. This causes the world to be updated in an **incorrect** order. First, due to the strategy followed to execute a **Moved items** transformation, the *Turtle* is left in the *Garden*, as “Kitchen” is not the player’s

5.2. Scenario A: Emma and her turtle

inventory nor the inventory of an NPC⁵. Second, the player’s location is changed to the *Kitchen*. Luckily, this usual error does not result in a confusing *narrated outcome*: as the player’s location changed, the *narrated outcomes* are not shown, and a *scene description* is generated instead (steps 2A and 2B of Figure 4.7 in Chapter 4).

There was only one PAYADOR error that was not related to the order in which the *transformations* are executed. It happened during turn #9 of TesterE’s playthrough. The user input was “agarro el otro martillo y me dirijo a la cocina” (*I take the other hammer and go to the Kitchen*), and the LLM suggested a **Moved items** transformation: “Moved object: <Un martillo verde> now is in <Inventory>” (“Un martillo verde” means *A Green hammer*). This does not seem problematic, until we take a look at the previous turn. In it (turn #8), TesterE said “miro el martillo verde, no será hojita?” (*I take a look at the green hammer, could it be hojita?*). Therefore, when in turn #9, the user input mentions “the other hammer”, it is a clear reference to the *Grey hammer*. This error is related to the lack of memory in the PAYADOR approach, and I will come back to it later.

Other interesting errors

Beyond **LLM-t** and **WTracker-t** errors, there are other errors that are interesting to empirically understand the limitations of our approach.

The first happens when the LLM responses break the narrative immersion. For instance, TesterA asked “Is Hojita, my pet, a turtle?” and the LLM answered (through a *narrated outcome*) “Hojita is not mentioned in the current world description. Therefore, I cannot answer your question about whether Hojita is a turtle.”. That is true: since the player is in the *Art studio*, the *fog of war* prevents the LLM to get information about the *Turtle*, that is in the *Garden*. In Appendix C.1 there is another interesting example in which TesterB asks “Do I see any doors?” and gets confused after the LLM answer, which lacks of common sense.

The second is error that sometimes, the LLM answers following a “Virtual assistant” style. For instance, TesterE asks Laura about the *Turtle*, and she answers: “Laura, en este momento, no parece tener a la tortuga. Puedes preguntarle directamente o buscar en otros lugares. ¿Necesitas que te ayude a buscar algo más?” (*It seems that Laura does not have the turtle at the moment. You can ask her directly or look for it in other places. Is there anything else I can help you with?*). Also, during turn #10, TesterF asks in the *Garden* “¿La tortuga que veo es Hojita?” (*Is Hojita the turtle I can see?*), and the LLM answers “No hay cambios en la ubicación de los objetos ni en los pasajes disponibles. Sí, la tortuga que ves en el jardín es Hojita” (*There are no changes neither in the location of the objects nor the available passages. Yes, the turtle you can see in the garden is Hojita*). Evidently, this not an immersive way to answer clarification questions.

The third aspect that may break the immersion is the lack of memory of PAYADOR. For instance, during turn # 7, TesterD goes from the *Kitchen* to the *Art studio*, and the *scene description* is “You’re still in the art studio, surrounded by the familiar sights and smells of oil paints and turpentine. (...)”. The word “still” confuse the player, so the answer is “i was not in the art studio! i was in the kitchen, looking for Hojita, remember?”. Unfortunately, I have to answer TesterD that no, PAYADOR does not remember that. It is interesting that, after that user input, the LLM suggests a **Player movement** transformation from the *Art studio* to the *Kitchen*; of course, that was

⁵As explained before, if the destination of the item movement is not recognized, it is left in the current player’s location.

Chapter 5. An Empirical Evaluation of PAYADOR

counted as an **LLM-t** error. Something similar happens to TesterG at the start of the session, when asking Laura to borrow her key. After getting the key in turn #2, TesterG started role-playing Emma, and having a conversation with her mom. This conversation lead to a situation in which Laura said she had no key to lend, what really confused the player. TesterG has experience playing TTRPGs before, so this example shows how this lack of memory restricts the narrative abilities of the modeled GM during the co-creation process. Since this is an illustrative example of this phenomenon, Appendix D contains the full translated example with some annotations.

The list of previous *scene descriptions* that PAYADOR maintains can also introduce errors, as I discussed in Section 4.5. For instance, in turn # 11, TesterG already had achieved *Milestone (II)*, but the *scene description* included the following sentence:

(...) la puerta que lleva al Jardín sigue bloqueada por un candado fuerte con el mismo escudo de armas que la llave que tienes en tu inventario

(...) the door that leads to the Garden keeps being locked by a strong lock engraved with the same coat of arms than the key you have in your inventory.

From a narrative point of view, this error is critical. Since the state of the world did not mention the *lock*, the only reason why the LLM decided to include it in the *scene description* is because it was present in a previous *scene description*. This error then led to a **LLM-t** error, since TesterG said “quiero abrir el candado al jardín” (*I want to open the lock (of the door) leading to the garden*), and the LLM suggested an **Unblocked locations** transformation, even when there were no blocked locations in the world state. Thanks to the way in how PAYADOR works — rendering the world state from scratch at the beginning of each turn — narrative errors like these can be quickly fixed through a clarification question, like: “Is the Garden door locked?”. However, they can really confuse players and make them have delays like those noticed in Figure 5.3.

Beyond critical cases like that one, another issue related to the *scene descriptions* is that they can be repetitive hence somewhat boring and not engaging, what usually happened when using Llama to narrate the scenes. I will come back to this in Section 5.4.

The beauty is in the details

Of course, not everything is bad, and there are many interesting emergent qualities that PAYADOR exhibited in these eight playthroughs.

The first, and most important one, is that PAYADOR has the gamemastering ability to answer clarification questions (as introduced in Chapter 4) based on the facts included in the component *descriptions*. For example, when TesterA said “Where can I go from the art studio?”, the LLM correctly answers (through a *narrated outcome*) “You are in the art studio. From here, you can go to the Kitchen.” Moreover, since the world state is rendered from scratch at the start of each turn (as explained in Section 4.5), every turn is a new opportunity to help *dialogue grounding* between the player and the system through these clarification questions. This was the case of TesterG, that experienced several errors caused by noisy *narrated outcomes*. After asking a clarification question, the fact that the *turtle* was not in the inventory was clear, allowing the player to continue the adventure.

5.3. Scenario B: Venancio looking for Artigas

The second observation is that the *previous visited locations* also had a positive effect. For instance, in turn #12, TesterB is back in the *Kitchen*, and the LLM describes the scene taking that into account: “You’re back in your kitchen. Sunlight streams through the windows. You’re still holding the grey and green hammers. The garden door remains locked. The art studio is accessible.”. The words “still”, “back” and “remains” gives the player a sensation of being in a solid and consistent fictional world. In general, it seems that Gemini performed better in these cases.

The third observation is related to the positive cases where PAYADOR could perform many transformations at once. For example, both TesterA and TesterF got the *grey hammer* and went to the *kitchen* at once, and TesterB even picked up both hammers at once. However, the most interesting example is that TesterG said “quiero agarrar el martillo pequeño” (*I want to take the small hammer*), and the LLM predicted a **Moved items** transformation to save the *Green hammer* in the player’s inventory, based on the *descriptions* of the *Green hammer*. Further research is needed to understand why the reference to “Hojita” caused many **LLM-t errors**, but references to the hammers like this one often worked like a charm.

The fourth and final observation is that LLMs seem to enhance the narrative experience when taking the *creative responsibility*, introducing details that make the world feel alive. For instance, TesterF says “Le pregunto a Laura si sabe dónde está Hojita.” (*I ask Laura to know if she knows where Hojita is*), and the LLM answers:

‘Le preguntas a Laura si sabe dónde está Hojita. Laura te responde: ‘No la he visto, cariño. ¿Has revisado bajo los muebles del taller?’

You ask Laura to know if she knows where Hojita is. Laura answers “I haven’t see her, darling. Have you checked under the Art studio furniture?”.

Even when there is nothing under the furniture, this LLM answer causes the player to engage in the objective, and in the next turn TesterF asked if Hojita was there. Naturally, this adds an additional narrative layer to the experience. Appendix D includes some turns taken from the playthrough of TesterD, during which the LLM took the creative responsibility many times.

5.3 Scenario B: Venancio looking for Artigas

The second scenario was designed with slightly different objectives. In this scenario, the player acts as Venancio, a *gaucho*⁶ **looking for Artigas**. Artigas is regarded as the most important national hero of Uruguay, and *gauchos* — along with other inhabitants of the South American region — were part of his military campaigns.

The focus here is on testing PAYADOR in a magical world, where historical facts and real-world physics are defied. We wanted to check how custom pre-authored worlds could be playable when the system had to take the *creative responsibility*. Accordingly, in this scenario, the player had to overcome two magical barriers that could not be solved by acting according to real-world physics.

In Figure 5.4 there is a graphical representation of this world. The *descriptions* of the ten components (2 **Character**, 4 **Item**, 3 **Location**, and 1 **Puzzle**) that are part of it, are as follows:

⁶According to the Real Academia Española dictionary: *Inhabitant of the pampas of South America, skilled in horsemanship and livestock work.*

Chapter 5. An Empirical Evaluation of PAYADOR

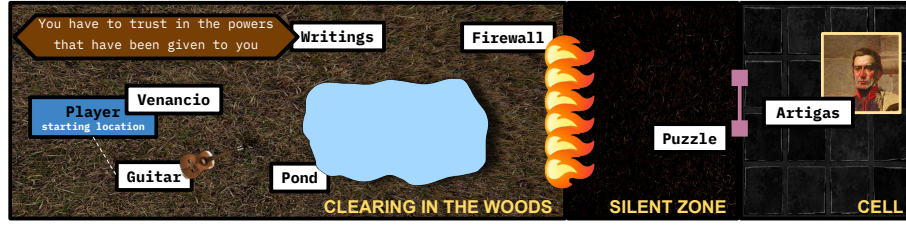


Figure 5.4: A graphical representation of Scenario B. The 10 components are represented: 2 characters (*Artigas*, an NPC, and *Venancio*, the player), 4 items (*Guitar*, *Pond*, *Writings*, and *Firewall*), 3 locations (*Clearing in the woods*, *Silent zone* and *Cell*), and a puzzle (*Puzzle*)

- **José Artigas:** [“Uruguay’s national hero”, “He is very weak after being locked up for so long.”]
- **Venancio:** [“An Uruguayan gaucho in his 40s”, “He belongs to the Artigas army”, “He has the magical power to summon a giant wave of water with which he can put out fires or moisten the ground.”]
- **Writings:** [“There is something written on the wall.”, “It says ‘You have to trust in the powers that have been given to you’.”]
- **Pond:** [“A pond full of crystal clear water”, “The water is so clear that it works like a mirror”]
- **Firewall**⁷: [“The flames are very hot”, “It’s 3 metres high”, “It is impossible to cross them, neither walking, nor running, nor jumping.”]
- **Guitar:** [“A classic guitar with 6 strings”, “It sounds great”]
- **Clearing in the woods:** [“A clearing in a eucalyptus forest near the Uruguay River”, “You can hear the sound of the animals that live in the trees of this forest.”]
- **Silent zone:** [“The forest continues in this part”, “Unlike the previous area, this area is very silent and not even the slightest sound can be heard.”]
- **Cell:** [“A square cell of two square meters”, “The interior is full of plants that grew outside”]
- **Puzzle:** [“A strong magic is generating an impassable wall”, “Magically, as you get closer, some blue letters appear explaining what the riddle to solve is.”].
Riddle: “You have to whisper the name of the river located on the southern coast of the Banda Oriental”; **Expected answer:** “Río de la Plata”.

Exactly like for Scenario A, these *descriptions* comprise all the details that PAYADOR knows about the world, and the testers did not read them before playing it. Also, the testers did not see this graphical representation in advance, since we designed it during the writing of this thesis.

5.3.1 Milestones and evaluation goals

To finish the scenario, the player has to find Artigas. Naturally, since the players do not know he is in the *Cell*, they had to explore the environment. In order to enter the *Cell*, the player must first put out the *firewall* and then solve the riddle in the *Silent*

⁷After running all the experiments we realized this component should be called *Wall of Fire*. Apparently, the word *firewall* generated no problems during the experiments in English.

5.3. Scenario B: Venancio looking for Artigas

zone, that will unlock the *Cell*. To put out the *firewall*, the player has to use *Venancio*’s magic to summon a giant wave of water. The **three milestones** we identify here are:

- **Milestone (I)**: Put out the firewall. This will unblock the passage to the *Silent zone*, where the *Puzzle* is located.
- **Milestone (II)**: Solve the riddle. Whispering “Río de la Plata” in the *Silent zone* will unblock the passage to the *Cell*, where *Artigas* is located.
- **Milestone (III)**: Find Artigas. Once the player is inside the *Cell* with *Artigas*, the scenario ends.

The evaluation goals of this scenario are very similar to those of Scenario A. The first evaluation goal was to test **PAYADOR features**. In this scenario there is no need of taking and leaving items, but there is a strong focus on the **Unblocked locations** transformation. Therefore, the goals here are more generic. Specifically, we wanted to:

- Check if the player can explore the locations without any issue
- Check if the system can successfully detect when the player achieved the objective, symbolically checked by the *World* class.
- Check if the system can correctly update the world when the *transformations* predicted by the LLM are correctly written.

Naturally, since the player has to trigger two **Unblocked locations** transformations to reach *Milestone (III)*, this scenario has a stronger focus on evaluating the **transformations predicted by LLMs** in a magical setting. Specifically, we wanted to:

- Check if the LLM can read the *writings* on the walls to the player, that suggest to “trust the powers that have been given to you”
- Check if *descriptions* can inform the suggested *transformations*. Specifically, check if the player can use *Venancio*’s spell of summoning a giant wave of water to trigger an **Unblocked locations** transformation. Analogously, check if the LLM can check the player’s answer to the riddle, suggesting an **Unblocked locations** transformation in case of a correct answer.
- Check if the LLM does not suggest an **Unblocked locations** transformation if the player does not use a reasonable strategy to do so.

Finally, as well as for Scenario A, we wanted to extract observations on the LLM ability to keep the **narrative immersion**:

- Check if the *narrated outcomes* of a wrongly predicted *transformation* may confuse players.
- Check if *scene descriptions* are good enough to let the player know what power *Venancio* has, and what is the *Puzzle* to open the *Cell*. Both facts are critical to finish the scenario.
- Check if the LLM can narrate events without breaking the immersion.

5.3.2 Quantitative results

The eight testers played this scenario after playing the previous one. The only exception was TesterH, that first played this scenario, and then played Scenario A. With this exception we wanted to check if the testers had to use more turns to finish the first scenario than the second one. I will come back to this soon.

Chapter 5. An Empirical Evaluation of PAYADOR

| Scenario B | Language | LLM | #Turn <i>Milestone (I)</i> | #Turn <i>Milestone (II)</i> | #Turn <i>Milestone (III)</i> |
|------------|----------|--------|-------------------------------|--------------------------------|---------------------------------|
| Gold | - | - | 1 | 3 | 4 |
| TesterA | English | Gemini | 3 | 6 | 7 |
| TesterB | English | Gemini | 9 | 12 | 13 |
| TesterC | English | Llama | 4 | 9 | 10 |
| TesterD | English | Llama | 6 | 8 | 9 |
| TesterE | Spanish | Gemini | 12 | 15 | 16 |
| TesterF | Spanish | Gemini | 2 | 5 | 18 |
| TesterG | Spanish | Llama | 5 | 9 | 10 |
| TesterH | Spanish | Llama | 1 | 12 | 13 |

Table 5.4: Number of turns to achieve each milestone by each of the testers. The second and third rows indicate which LLM and what language was used for that experiment.

Like for Scenario A, all the testers could finish the scenario, and two English examples of these playthroughs are included in Appendix C.2. This time, TesterH had to repeat the experiment after being confused about the actual world state narrated by the LLM.

Turns to reach each milestone

Table 5.4 shows the summary of the playthrough of each tester, based on the turn they reached the milestones in. This time, the “Gold” playthrough is based on the following sequence of **single transformations**: (1) Put out the flames, (2) Go to the *Silent zone*, (3) Solve the riddle (*Puzzle*), and (4) Go to the *Cell*.

Comparing the results of Scenario A in Table 5.2 with the results of Scenario B in Table 5.4, it is clear that almost all testers could finish the second scenario in less turns. This is also the case for TesterH, that played Scenario B first. TesterF finished Scenario A really early and was in track to make it again, but during Scenario B had critical problems that caused a huge delay. This can be better visualized in Figure 5.5, where TesterF is the first to achieve *Milestone (II)* but the last one to achieve *Milestone (III)* one. I will come back to this case later.

Errors during the playthrough

Just like we did to analyses the playthroughs of Scenario A, we repeated the manual annotation process for this scenario, trying to find the same two type of errors: **LLM-t** and **WTracker-t**. The results of this manual annotation are available in Table 5.5. As can be seen, we did not find any **WTracker-t** errors in this playthroughs, and I will come back on this in the next section.

Now talking about specific cases, TesterH was exceptionally affected by errors. As can be seen in Figure 5.5, the player had a notable delay after achieving *Milestone (I)*. However, only one *transformation* error was found during that playthrough. This shows how critical other errors that are not related with *transformations* (hence not annotated and not reported in the table) can be. The main problem here was a sequence of wrong *scene descriptions* that confused the player. The case of TesterF was similar, being affected by *transformation* errors but also by other type of errors: after achieving *Milestone (II)*, the player needed 13 more turns after a wrong and confusing answer to a clarification question. In an opposite way, TesterE — that had the same number of LLM errors as TesterF — was somewhat helped by those errors (i.e. the errors helped TesterE to achieve some milestones). I will come back to this in the next section.

5.3. Scenario B: Venancio looking for Artigas

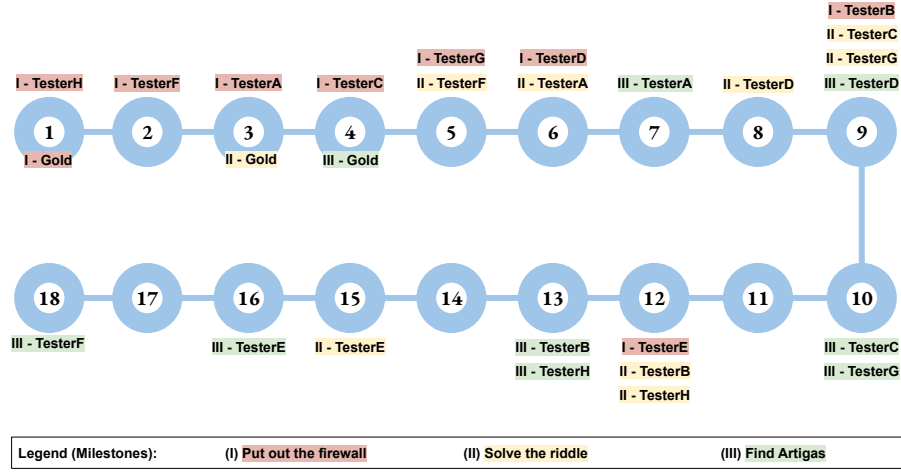


Figure 5.5: A graphical representation of the eight testers achieving the milestones while playing Scenario B. Each circle represents a turn, and they go from 1 to 18, the latest turn when a tester finished the scenario.

| Scenario B | Error name | Errors in <i>Milestone (I)</i> | Errors in <i>Milestone (II)</i> | Errors in <i>Milestone (III)</i> | Errors during the whole test |
|--------------------|------------|-----------------------------------|------------------------------------|-------------------------------------|---------------------------------|
| TesterA | LLM-t | 0 | 1 | 0 | 1 |
| (Gemini — English) | WTracker-t | 0 | 0 | 0 | 0 |
| TesterB | LLM-t | 1 | 0 | 0 | 1 |
| (Gemini — English) | WTracker-t | 0 | 0 | 0 | 0 |
| TesterC | LLM-t | 0 | 1 | 0 | 1 |
| (Llama — English) | WTracker-t | 0 | 0 | 0 | 0 |
| TesterD | LLM-t | 0 | 0 | 0 | 0 |
| (Llama — English) | WTracker-t | 0 | 0 | 0 | 0 |
| Total | LLM-t | 1 | 2 | 0 | 3 |
| (English) | WTracker-t | 0 | 0 | 0 | 0 |
| TesterE | LLM-t | 2 | 2 | 0 | 4 |
| (Gemini — Spanish) | WTracker-t | 0 | 0 | 0 | 0 |
| TesterF | LLM-t | 0 | 0 | 3 | 3 |
| (Gemini — Spanish) | WTracker-t | 0 | 0 | 0 | 0 |
| TesterG | LLM-t | 1 | 0 | 0 | 1 |
| (Llama — Spanish) | WTracker-t | 0 | 0 | 0 | 0 |
| TesterH | LLM-t | 0 | 1 | 0 | 1 |
| (Llama — Spanish) | WTracker-t | 0 | 0 | 0 | 0 |
| Total | LLM-t | 3 | 3 | 3 | 9 |
| (Spanish) | WTracker-t | 0 | 0 | 0 | 0 |

Table 5.5: Number of errors for each type during the playthroughs of the eight testers playing Scenario B. The columns that count errors in a milestone consider the errors found in turns **after** the last milestone and in turns **before or during** achieving that milestone.

5.3.3 Qualitative observations

Following an identical structure as for Scenario A, in this section I will present the observations extracted after analyzing the **LLM-t** and **WTracker-t** errors, and then I will include other general observations.

Chapter 5. An Empirical Evaluation of PAYADOR

LLM-t errors

I will start by analyzing what happened in the playthrough of TesterF, that caused an unexpected delay. As I previously mentioned, the player reached *Milestone(II)* in turn #5, and only needed to enter the *Cell* to reach *Milestone(III)* in the next turn. But then, a subtle detail created a huge confusion. After unlocking the door to the *Cell*, the player asked “¿Qué hay en la celda?” (*What is inside the cell?*), and the LLM answered “En la celda no hay nada. Es una celda vacía y fría.” (*There is nothing inside the cell. It is an empty and cold cell*). The original cause of this error is the *fog of war* rule to render the world state, but the LLM should have answered in a more discrete way, encouraging the player to explore.

As TesterF understood that the *Cell* was of no interest, then started exploring the *Silent zone* and the *Clearing in the woods* again. In that context, three LLM errors occurred. Both in turn #9 and turn #15, the player tried to go to the *Silent zone* by writing its name with two capital letters (**Zona Silenciosa**). Since the LLM did not include the correct name of the *location* in the suggested *transformation* (“Your location changed: <Zona Silenciosa>”), the consistency checks prevented that movement because the string did not match with any *reachable location* (i.e. the checks are **case-sensitive**). As a consequence, the *narrated outcomes* told the player that the new location was “Zona Silenciosa”, what caused even more confusion. Another error occurred in turn #17, when the player finally got to the *Silent zone* again. Although the *cell* was already open since turn #5, the LLM had an error when describing the scene, saying that the *cell* was locked: “Estás en la Zona Silenciosa. El silencio es total. Puedes ir al Claro en el monte. La celda sigue cerrada, con las mismas letras azules cerca. Tienes tu guitarra” (*You are in the Silent zone. There is absolute silence. You can go to the Clearing in the woods. The cell is still locked, with the same blue letters. You have your guitar*). Confused, TesterF said “Susurro ‘Río de la Plata’ para volver a abrir la celda.” (*I whisper ‘Río de la Plata’ to open the cell again*), and the LLM suggested an **Unblocked locations** transformation “Blocked passages now available: <Zona silenciosa>”. The error not only lies in failing to understand what the player wanted to say (to unblock the *Cell*, not the *Silent zone*), but also suggesting a *transformation* to unblock the passage to the *Silent zone*, an already reachable location. Worse, the LLM answered (through *narrated outcomes*) “Susurraste ‘Río de la Plata’ pero nada sucede. La celda permanece cerrada. Sigues en la Zona Silenciosa.” (*You whispered ‘Río de la Plata’ but nothing happens. The cell keeps locked. You are still in the Silent zone*). In that moment, TesterF asked for help to finish the scenario.

In the playthroughs of TesterG and TesterH there were similar LLM errors, too. TesterG said “camino hacia el estanque” (*I walk towards the pond*), and the LLM suggested “Your location changed: <Estanque>”, what is incorrect since “Estanque” (*Pond*) is not a *location*. Similarly, TesterH said “Volver al puzzle mágico” (*Go back to the magical puzzle*), and the LLM suggested “Your location changed: <Puzzle mágico>”, what is incorrect for exactly the same reason. In both cases PAYADOR filtered out the wrongly suggested transformations.

Another strange error happened during turn #6 of TesterB. Convinced that the guitar was needed to summon the giant wave of water, the player said “Play a tune in the guitar and try to summon a giant wave from the lake to put out the fire with a song, while I sing I trust in the powers that have been given to you (music)”. The LLM answered “You play a tune on your guitar. The lyrics of your song mention trusting in the powers given to you, as the writings on the wall suggested. However, nothing happens. The fire remains and blocks the passage to the Silent Zone. The magical powers seem to require a different trigger or a more precise invocation. ”. What is strange, is that almost the same meaning was conveyed by the player utterance of turn #9: “I believe that the powers that have been given to me is the ability to create

5.3. Scenario B: Venancio looking for Artigas

music, thus, I try to use this powers to put out the fire wall by summoning a giant wave from the lake” (sic). After that input, the LLM suggested “Blocked passages now available: <Silent zone>” and the player achieved *Milestone (II)*

One of the most worrying errors happened when the players wanted to unblock locations, and the case of TesterE is very illustrative. It all started in turn #11, when the player creatively says “agarro mi cantimplora y la lleno con agua” (*I take my canteen to fill it with water*). This is really interesting, since there was no canteen previously mentioned in the world state, but it really makes sense that a *Gaucha* had some way of carrying water. It is also important to clarify that although TesterE never asked about Venancio’s powers, the LLM never mentioned anything about it, not even in the starting narration of turn #1. As a consequence of that user input an **LLM-t** error occurs, since it suggests a **Moved items** transformation: “Moved object: <cantimplora> now is in <Inventory>”. Naturally, since “cantimplora” (*canteen*) was not an item in the world, the consistency checks prevent this from happening. However, the *narrated outcomes* the LLM generated as a consequence, confused the player: “Agarraste tu cantimplora y la llenaste con agua del estanque. Ahora la llevas en tu inventario.” (*You took your canteen and filled it with water from the pond. Now you carry it in your inventory*). Naturally, thinking that the *canteen* is in the inventory, TesterE tries to do something with it: “con sigilo me aproximo al fuego, vacio mi cantimplora sobre él” (*stealthy, I get closer to the firewall, and empty the canteen over it*). The LLM then suggested this **Unblocked locations** transformation “Blocked passages now available: <Zona silenciosa>” and the *narrated outcomes* are “Con sigilo te aproximas al fuego, vacías tu cantimplora sobre él. El agua no es suficiente para apagar las llamas, pero sí lo suficiente para generar vapor que permite ver un pasaje tras el muro de fuego, ahora accesible. La Zona silenciosa ahora está disponible”. (*Stealthy, you get closer to the fire, empty your canteen over it. The water is not enough to put out the flames, but it is enough to generate steam that allows to see a passage behind the firewall, now accessible. The Silent zone is now available.*)

Some turns later, TesterH reaches *Milestone (II)* by saying “miro el puzzle, hago fuerza de mi y resuelvo el puzzle” (*I take a look at the puzzle, I focus, and then I solve the puzzle*). As a consequence, the LLM suggests an **Unblocked locations** transformation: “Blocked passages now available: <Celda>”. A similar error occurred during the playthroughs of TesterA and TesterC, when they unlocked the *Cell* door by saying a wrong answer. Instead of answering “Rio de la Plata”, they answered “Rio Uruguay”, which is another important river in Uruguay, but not located in the south of the region. All these cases are strongly related with the observations we did after the experiments I discussed in Chapter 3: LLMs usually try to adjust their outputs to the prompt. Since challenging obstacles are at the core of TTRPGs, and they should be consistent and hard to overcome, these errors are critical and worth further research.

WTracker-t errors

As I previously mentioned, in this scenario there were no **WTracker-t** errors. This is probably related to the fact that most of the **WTracker-t** errors in Scenario A were caused by the order in which the *transformations* are executed (**Moved items**, then **Unblocked locations**, and then **Player movement**). However, this scenario is strongly focused on the **Unblocked locations** and the **Player movement** transformations, which rarely need to be executed in a different order.

Since the players get a *scene description* **after** moving to a new *location*, it is not likely that they take actions that need a **Player movement** and a **Unblocked locations** transformation, in that order (because they do not know what is there before moving). For instance, after achieving *Milestone (I)*, an input like “I go to the

Chapter 5. An Empirical Evaluation of PAYADOR

Silent zone and then whisper ‘Rio de la Plata’ to unlock the *Cell*” is not probable at all, since the player does not know there is a riddle there.

Other interesting errors

Previously, during the analysis of the quantitative results, we observed that TesterF and TesterH had abnormal delays by looking at how separated their milestones were. While the errors during the playthrough of TesterF were a combination of **LLM-t** errors, wrong *narrated outcomes* and wrong *scene descriptions*, the case of TesterH is different, as in that playthrough we only found one **LLM-t** error. The main problem in that case were the *scene descriptions*, as the first time the player entered the *Silent zone* the LLM did not explain the riddle. Although TesterH asked about the riddle many — many — times, the LLM did not explained how to solve it. As a consequence, TesterH went back to the *Clearing in the woods* to check if some clues had been missed. After some turns, back in the *Silent zone*, TesterH said again “Volver al puzzle mágico” (*Go back to the magic puzzle*) and the LLM answered:

Regresas al puzzle mágico, donde debes resolver el acertijo para desbloquear el pasaje a la Celda. Recuerda que debes susurrar el nombre del río que baña la costa sur de la Banda Oriental.

You return to the magic puzzle, where you must solve the riddle to unlock the passage to the Cell. Remember you have to whisper the name of the river located on the southern coast of the Banda Oriental.

Naturally, this is one of the risks of integrating LLMs in games: when they fail, they can make the player feel stuck, confused or frustrated. Since this is an illustrative case, it is included in Appendix D.

Like for the first scenario, the LLM again broke the narrative immersion. For instance, during the communication chaos TesterH suffered, many clarification questions were answered with “Nothing happens”. That is similar to what happened to TesterB, when asked “Am I dreaming?”, as the player had the sensation of being in a dreamlike setting. A similar error affected TesterE, when tried to play a song to call Artigas, and the LLM answer included an unexpected question to the player:

Las pinturas dicen ‘Hay que confiar en los poderes que se nos han otorgado’. ¿Te refieres a tus poderes mágicos?

The writings say ‘You have to trust in the powers that have been given to you.’ Do you mean your magical powers?

It seems like the LLM mixed up the user input with the rest of information in the prompt: the player has no incidence in those magical powers as they are part of the *descriptions* of *Venancio*.

In this scenario the *previous scene descriptions* also caused the LLM to mess up some details. For instance, TesterF was back in the *Clearing in the woods* after putting out the fire, but the *scene description* said the *firewall* was preventing the passage to the *Silent Zone*. We checked the symbolic world state and, indeed, the wall of fire is not mentioned there, so the only way it could have been considered by the LLM was due to past *scene descriptions*. Consequently, the player answered “El muro de fuego no debería estar, lo apagué con mis poderes” (*The wall of fire should not be there. I put it out using my powers*) and the LLM said: “Aparentemente, tus poderes mágicos han extinguido el muro de fuego. Sin embargo, no hay ningún muro de fuego

5.3. Scenario B: Venancio looking for Artigas

visible en el Claro en el monte.” (*Apparently, your magic powers have extinguished the wall of fire. However, there is no wall of fire visible in the Clearing in the woods*). The comment “Apparently, your magic powers have extinguished the wall of fire” is unnecessary, as it contradicts the sentence succeeding it “However, there is no wall of fire visible in the Clearing in the woods.” This seems to unveil again the observations reported in Chapter 3: the LLM tries, as much as possible, to include the user input in the output. Further work should focus on this kind of problem, as it totally breaks the narrative immersion.

Finally, a very interesting semantic problem occurred when TesterE asked a rhetorical question: “miro al cielo y me pregunto: podré levantar el estanque?” (*I look at the sky and I wonder, would I be capable of lifting the pond?*). As humans, we perfectly understand that this player was referring to “making the water flow up”, and “using it to put out the wall of fire”. However, the LLM understood that the player **literally** wanted to lift the pond, as if it were a heavy object. It seems that, as of 2025, not all linguistic problems are solved by LLMs.

The magic of Role-playing

Just like happened in the previous scenario, there were many cases where players engaged in role-playing.

Some players tried to make *Venancio* sing. TesterC tried to play a song to call Artigas, and the LLM continued by answering: “You start playing the melody on your guitar, the sound echoes through the forest, but there’s no response from Artigas yet. The writings on the wall seem to be encouraging you to trust in your powers. The pond’s calm surface reflects the trees surrounding the clearing.” TesterE tried the same, but including verses from “A Don José”, a real song that refers to José Artigas.

Other players tried to use the *Guitar* to summon the wave of water. Probably they came up with this idea since in *Dungeons & Dragons* there is a *class* — the *Bard* — who summon spells using musical instruments. TesterG tried this by saying: “toco la melodía de llamado del agua” (sic) (*I play the melody of the call of the water*) and, as a consequence, achieved *Milestone (II)*. TesterD also tried this in turn #4, as can be seen in the playthrough in Appendix C.2.

A related detail that we found interesting is that, when trying to unlock the *Cell*, TesterG tried to sing a song about “Río de la Plata”. As a consequence the LLM did not suggest an **Unblocked locations** transformation. However, in the next turn, the player said “susurro ‘Río de la Plata’ ” (*I whisper ‘Río de la Plata’*) and then the LLM suggested to unlock the *Cell*. This seems to indicate that the LLM could tell that *singing* is definitely not *whispering*. Although the LLM was a bit inconsistent with this kind of considerations, we celebrate that the system can handle magical and creative solutions from the player in some cases. As I explained in Section 4.2, this is the main reason why we propose *transformations*: to embrace the creativity of the players.

Role-playing dialogues between the players and the LLM were one of the most eye-catching PAYADOR characteristics during the experiments. However, as also happened during Scenario A, there were other — more common — narrative details that worth a look, such as the ability to answer clarification questions about the environment. For instance, TesterF asked “¿Cuáles son mis poderes?” (*What are my powers?*) and the LLM correctly answered:

Tus poderes como Venancio, gaucho uruguayo de los soldados de Artigas, incluyen la habilidad de invocar una ola gigante de agua. Puedes usarla

Chapter 5. An Empirical Evaluation of PAYADOR

para apagar fuegos o humedecer la tierra. Las pinturas dicen ‘Hay que confiar en los poderes que se nos han otorgado’, lo cual parece una confirmación de tus habilidades.

Your powers as Venancio, a Uruguayan gaucho of Artigas’ army, include the ability to summon a giant wave of water. You can use it to put out fires or moisten the earth. The paintings say ‘We must trust in the powers that have been given to us’, which seems like a confirmation of your abilities..

TesterA asked the same question and the LLM answer can be seen in Appendix C.2.

Other examples of good answers to clarification questions are those utterances that explain the riddle to the players. For instance, TesterF asked “¿Qué dicen las letras azules?” (*What do the blue letters say?*) and the LLM correctly answered:

“Las letras azules dicen: ‘Hay que susurrar el nombre del río que baña la costa sur de la Banda Oriental’. Para desbloquear el pasaje a la Celda, tendrás que resolver este acertijo.”

The blue letters say: ‘You must whisper the name of the river located on the southern coast of the Banda Oriental’. To unlock the passage to the Cell, you will have to solve this riddle.

Finally, TesterH got a beautiful narrative detail during the communication chaos (available in Appendix D). When shouting in the *Silent zone* in Turn #4, the LLM answer was:

No hay respuesta a tu grito. El silencio es absoluto en esta zona. Continúas sin escuchar nada

There is no answer to your call. The silence is absolute in this area. You continue without hearing anything.

This answer mixes both facts, *Venancio* calling *Artigas* and the absolute silence that characterizes that place. This is great, since it gives the players the feeling that every detail in the fictional world is in harmony: the silence is absolute, no mattering who or what is the one that tries to generate the sound.

5.4 Overall findings and future work

In the previous sections I detailed the experiments we carried out, with eight voluntary testers playing both Scenario A and Scenario B, in English or Spanish. I also discussed the quantitative results we obtained — framed by the definition of three *milestones* that represented the player’s progress in the scenario — and the qualitative observations we extracted. The general discussion was done with a focus on the *transformations*, for which we defined two errors: **LLM-t** (when the LLM makes errors while suggesting *transformations*), and **WTracker-t** (when the LLM was not involved in the mistake, caused by a design limitation of the PAYADOR approach). It is important to clarify that these two type of errors were only proposed to organize the discussion, i.e. distinguishing between the errors that are introduced by the LLMs, and those that can be fixed by working on the PAYADOR approach itself. However, regardless of the type of error, what is more important is to understand them, hence empirically defining the virtues and defects of PAYADOR.

5.4. Overall findings and future work

After finishing the two scenarios, we asked the participants their general opinions about the experience. We also asked them four extra questions: “Did you find any inconsistencies in the narratives?”, “Were there any errors that made it difficult for you to complete the scenario?”, “How *natural* do you consider the narratives were?”, and “Do you have any further comments?”.

In this final section I will make a general summary of the observations we did, clustering the virtues and defects we detected and also those the players commented.

5.4.1 Transformations

Let’s start by the most important aspect of our approach: *transformations*. The quantitative results show that all testers could achieve *Milestone (III)* in both scenarios, what seems to indicate that *transformations* are a viable approach. This is possible thanks to the current capabilities of LLMs — of working as *common-sense reasoning* oracle —, which are critical to tell what are the effects of **any** action the player can think of. However, there are two main problems that hinder better performance of the general GM model.

What is the order they should be executed in?

The first problem is the appropriate order the *transformations* must be executed in. In this first version of PAYADOR, we modeled three *transformations*: **Moved items**, **Unblocked locations** and **Player movement**. During and after the experiments, we noted that players quickly got immersed in the game. That was definitively good news, but it also meant that they were going to try **anything** they could think of, hence needing to trigger *transformations* in any order.

Thinking about it in a more theoretical way, this is totally reasonable. The players may want to:

- take an *item* in the current *location* and move to a new *location*. This implies: **Moved items** → **Player movement**.
- move to a *location* they already know, and take an *item* they also know is available. This implies: **Player movement** → **Moved items**.
- unblock a *location*, and move there. This implies: **Unblocked locations** → **Player movement**.
- move to a *location* they already know, and unblock a *location* potentially reachable from there. This implies: **Player movement** → **Unblocked locations**.
- unblock a *location*, and instead of moving there, take an *item* available in the current *location*. This implies: **Unblocked locations** → **Moved items**.
- take an *item* in the current *location*, and then unblock a new *location*. This implies: **Moved items** → **Unblocked locations**.

And those are only the combination of two *transformations*. When considering the three *transformations*, it gets worse. And even worse, if other *transformations* are added, such as *Health Points increased*, *Mood change*, or many other possibilities that system designers could consider.

The same problem arises when chaining several actions in classic Interactive Fiction games (Reed, 2023). However, in that case the world is closed, so several specific strategies can be designed to disambiguate the order in which the actions had to be executed. This is not the case when working towards Open-world Interactive Narrative: when the player agency is at its maximum, no one can predict what will be the context in which actions will be taken (i.e. what will the preconditions and contexts be).

Chapter 5. An Empirical Evaluation of PAYADOR

One of the possibilities is to incorporate planning strategies (neural, symbolic or neuro-symbolic) that, based on the current state of the world, suggest what is the order in which the detected *transformations* have to be executed. Another way of doing this would be by prompting the LLM to do so: say **what** *transformations* to apply, and also in **which** order they have to be applied.

The theoretical analysis may sound terrifying. However, from a pragmatic point of view, the players may not use **all** the *transformations* at once, and some combinations may not be needed. Therefore, some heuristics could be designed for the specific set of considered *transformations*, and further research needs to be done in this line.

LLMs may fail

The previous discussion did not take into account a potential failure of the LLM when predicting the *transformations*. In the vast majority of the playthroughs, the LLM made at least one error when suggesting *transformations*.

Sometimes, the errors were related to not calling the components by the name they have in the world state. But the most worrying error is when the LLM tries to incorporate the user input in the output. I first introduced this error in Chapter 3, when analyzing the performance of LLMs as out-of-the-box GM models. This is something that happens very frequently, and makes us reflect about how suitable are current LLMs to allow the players to play a challenging role-playing game. If “I solve the riddle” or “I take the bazooka” (see Figure 4.8 in Chapter 4) make the LLM immediately suggest a *transformation*, the players will never face an engaging challenge. Having these LLMs grounded to a structured representation allow to run consistency checks, as PAYADOR does, like the one shown in the *bazooka* example. However, as these narratives will take place in open worlds, the consistency checks cannot be more than those in harmony with the considered *transformations*. In other words, Open-world interactive storytelling demands general consistency checks, and not those classic in Closed-world systems that can be preprogrammed to check a specific answer, a specific action, or a specific use of a specific item (e.g. if it is the first time using *flashlight*₂₃, then reveal all the monsters in the room; else, the flashlight explodes and hurts the character holding it).

One of the possibilities to tackle this problem is to use Self-consistency (Wang et al., 2023) or an analogous method, where the same LLM — or other LLMs — can judge the decision of suggesting what *transformations* have to be done. Another possibility, not mutually exclusive with the previous one, is to have one LLM call per *transformation*. That way, the LLM can focus only on predicting the occurrence — or not — of a specific effect after the user input. Additionally, the instructions included in the prompts can be specific about those cases and, why not, the LLM can be fine-tuned to increase the performance even more.

We did not try other strategies (different to using a single call to the LLM) since, as I previously mentioned, we had two restrictions in mind: test the core ideas of the PAYADOR approach without further interference, and use as few LLM calls as possible. Naturally, other strategies could be tested in future extensions of this work.

5.4.2 Fog of war

Sometimes the LLM fails because it has not enough information. This was observed in some **LLM-t** errors, where the *fog of war* applied to the rendered world state filtered out details that the LLM would need to answer the player’s question, or predict the correct *transformation*. Although these errors could be also considered **WTracker-t** errors, as the *fog of war* strategy is part of the PAYADOR approach, we opted to

5.4. Overall findings and future work

consider them **LLM-t** errors, as the LLM should constrain its outputs to the available information.

And that is a part of the problem. The LLM tends to add information to generate a more complete answer and, if it does not have the information, it will probably hallucinate the missing details. This was exactly what happened when TesterF wanted to know what there was inside the cell, and the LLM answered that there was nothing there. After the experiment, in the evaluation questions, the tester confirmed that this was very confusing. In other cases, the LLM did not include new details, but answered in a non-immersive way, such as “you just asked a question about placing a non-existent ‘Turtle’ in the kitchen”. This was also noted by the players that experienced it.

A human GM would never do strong comments if at the moment does not have precise information about something that will be defined later. However, just to have another opinion, I asked a *Dungeon Master* what he would have done in the situation where he knew that the *Cell* was reachable from the *Silent zone*, and that *Artigas* was inside. He answered that he would said “The room is very dimly lit, but through the darkness you can see what appears to be a human silhouette at the back of the room” to encourage the player to explore (what did not do the LLM with TesterF). In other words, what my he was telling me is that the GM should have know that there was *Artigas*. Analogously, in Scenario A, the GM should have know where the *Turtle* was.

Therefore, we think that, as a strategy to prevent future inconsistencies, the *fog of war* strategy should be modified. It is not useless but, in some cases, it does not really model how GMs think about the fictional world. The *fog of war* restriction could be used to make the LLM focus during the world-update process, but also it could be disabled when the player asks a general question. Naturally, another option is to expand the amount of information the *fog of war* admits (e.g. including also the descriptions of *items* and *characters* that are in reachable locations), but without revealing information and spoiling the final of the scenario. Of course, further research is needed to decide what alternative is the best.

5.4.3 Lack of memory

Another error related with the design of our approach was its *lack of memory*. This characteristic of PAYADOR was decided taking into consideration the observations we extracted after the experiments I detailed in Chapter 3. As we saw that LLMs were prone to mix up details during gameplay, we decided that the *only memory* of PAYADOR had to be the structured world state. Although we were sure about that, we then created the *previous scene descriptions* list of strings to overcome the problem of repetitive scene descriptions. Unfortunately, we found again how problematic can this be, as the LLM mentioned details that were not part of the current state of the world.

We first think that this was caused by the fact that the LLM could access the *previous scene descriptions*, but not their *associated world states*. However, the LLM has access to the **current** world state, what indicates that it had enough information to avoid including inconsistent details. This was mentioned by several testers in their answers to the evaluation questions.

Naturally, those two *memories* (structured world state and *previous scene descriptions*) are not enough: the player naturally tries to make references to words that said before, or things that happened before. That is why we considered the *lack of memory* as a **WTracker-t** error: if the LLM cannot suggest the correct *transformation* when the player is mentioning something from the past, the error is PAYADOR’s fault.

All these observations take us to a point where we think that LLMs are not yet

Chapter 5. An Empirical Evaluation of PAYADOR

reliable to be the out-of-the-box gamemastering models but, at the same time, we know that the structured world state is not enough to get a full gamemastering model. There may be some terrain to explore in the middle of both extremes; a balance between structured representations and the list of strings. Also, a symbolic log could represent the events that actually happened after the consistency checks, in order to prevent saving noisy information (e.g. the player takes a non-existent *bazooka* and it is saved in the log, although the *bazooka* is actually never added for being filtered out by the consistency checks). Further research on how to correctly represent this log should be done in the future.

5.4.4 Narrated outcomes and scene descriptions

Narrating the scenes and the outcomes of the players' actions are one of the most important gamemastering tasks. Luckily, the simple strategy we followed, using the *narrated outcomes* and the *scene descriptions*, was good enough to help the human evaluators reach the three milestones and finish the scenario. However, some narrations were not good enough and others were very misleading, creating a strong confusion in some players.

The most problematic aspect was that the *narrated outcomes* and the predicted *transformations* were generated **at the same time**. When the *transformations* are correct, the *narrated outcomes* are really useful as a feedback for the player. However, when the *transformations* are rejected by the consistency checks, the *narrated outcomes* include details of events that did not actually happened according to the world state. This was very confusing to the players that, after reading a reasonable *narrated outcome*, they thought the fictional world was in a state that was not. This error caused many delays in their progresses, having to go back and forth between the different *locations* to understand what happened, where the *items* actually were, and so on. Naturally, this was mentioned by many players in the evaluation questions and we also consider it is the most important issue to fix in the PAYADOR approach.

Although the impact of these errors was noticeable, the fact that the world state was rendered from scratch each time, at the beginning of each turn, helped the players understand what was the actual state of the world, regardless of past narrations. In future experiments, we consider it would be really useful that the players have a “narrate again” option, that forces the system to render the world again, call the LLM again, and show a new narration instead, to allow easier *dialogue grounding*. Unfortunately, this “narrate again” option will not be a guarantee of an exhaustive description of the environment, since many times the *scene descriptions* not always mentioned all the available things. The most emblematic case is that of TesterH, who in Scenario B entered the *Silent zone* and the LLM did not mention what riddle needed to be solved in order to open the cell. After asking about it several times, the LLM did not answer the question, making the player wonder if the actual question of the *riddle* was also part of the *puzzle* itself. This was confirmed by the player in the evaluation questions.

There are many changes to be done. The first, and most important, is that the *narrated outcomes* should not be generated at the same time than the predicted *transformations*, and had to be generated **after** the consistency checks. The second, and very related to it, is that probably the *scene descriptions* and the *narrated outcomes* should be generated by a single module. Since the PAYADOR approach (see Figure 4.7 in Chapter 4) discards the *narrated outcomes* when the player is in a new location, sometimes the testers did not get feedback about what happened in the previous location. For example, if the player is in the *Garden* of Scenario A, and says “I take the turtle and pet her. I feed her with the lettuce leaf I have, and then, with her in my

5.4. Overall findings and future work

| | Language | Gemini | Llama |
|-----------------------|----------|--------|--------|
| <i>Minimum length</i> | English | 33 | 49 |
| <i>Average length</i> | English | 184.50 | 405.10 |
| <i>Maximum length</i> | English | 467 | 1620 |
| <i>Minimum length</i> | Spanish | 55 | 69 |
| <i>Average length</i> | Spanish | 189.17 | 279.40 |
| <i>Maximum length</i> | Spanish | 528 | 870 |

Table 5.6: Minimum, average, and maximum character lengths for the GM utterances generated by Gemini and Llama.

hands, I go to the *Kitchen*”, PAYADOR would discard the *narrated outcomes* about what happened in the *Garden*, and just describe what the player sees in *Kitchen*. Of course, this was not intended when designing the approach, and it was detected during the experiments. It is also important to highlight that this problem only occurred when then the player caused more than one *transformation*: if there was only **Moved items** or **Unblocked locations** transformations, the *narrated outcomes* were used; and if there was only a **Player movement** transformation, the *narrated outcomes* would be discarded to show a more general *scene description*.

Currently, the LLM gets a template-generated rendered world state as part of the prompt to describe the scene. However, the LLM always gets the same information, and it is not previously contextually filtered or prioritized. Therefore, another interesting option to explore is to have different *scene description* prompts for the LLM, based on the current state of the world. For instance, if there is a blocked passage, generate an additional instruction to emphasize the importance to describe it, such as “The passage to <location> is locked by <component>, as mentioned in the world state. Please, include this detail in your output, as it is of critical importance”. Or, for example, if the player has an *Item* or it is near a *location* related to the objective, the LLM could get instructions such as “The player is close to achieving the objective. Please, emphasize this situation, and remind the player what the objective is”. This contextual-aware world state rendering would certainly open many possibilities.

A fourth aspect to be further analyzed is that the LLMs may sound very repetitive, what was commented by most of the players. Even having the *previous scene descriptions* list as an input, some details were repeated timer after time. This problem may be quickly tackled following the previously approach of generating better instructions for the LLM based on the world state (i.e. prioritizing some components when generating the scene description). Additionally, we noted that the GM utterances (*scene descriptions* and *narrated outcomes*) generated by Llama seemed to be so much longer than those generated by Gemini. To check the factuality of this intuition, we calculated the minimum, average, and maximum **character length** of all the GM utterances, both in Spanish and English, and the results are included in Table 5.6. The results show that, indeed, the Llama narrations were longer. However, a deeper linguistic analysis is needed to tell if they were, in the end, more verbose than those generated by Gemini. In Appendix C there are two examples of playthroughs using Gemini, and two using Llama.

All these ideas are not mutually exclusive, and can be integrated in a **single module** that uses different prompts — according to the state of the world — in order to generate **all** the GM utterances. Having in mind the consistency problems that usually text-based GM models face, a module like this would also make PAYADOR better take advantage of the *neuro-symbolic truth* saved in the world state. Scene descriptions, clarification questions, NPCs utterances, narrated outcomes; everything generated by a single context-aware module that takes into consideration the facts

Chapter 5. An Empirical Evaluation of PAYADOR

represented in the world state. Since the GM may need to have a general vision of the fictional world and remember events that happened before, the problems to be faced when building this module are closely related to the previously discussed *lack of memory* and *fog of war* problems. Therefore, this module would have to consider contextual information about the world state, the *user input*, the *log of events*, the recent *transformations* and other aspects that may be relevant.

There is no doubt: the gamemastering skill of exhaustively describing the scene without ambiguity and verbosity, while also being mysterious about details that the player should later discover, is definitively one of the main research lines that need to be approached in the future.

5.4.5 PAYADOR and LLMs in the co-creation process

Finally, one of the most interesting observations was that the PAYADOR approach allows narrative co-creation between the player and the computer. When having to take the *creative responsibility*, the LLMs can generate interesting details that make the world feel alive. It feels special when they consider the *user input* to make creative comments, as usually human GMs do to co-create the fictional world with the players. One of the examples I discussed was that in which the tester played some music with the guitar in the *Silent zone*, but it could not be heard. That was not a physical law mentioned in the original *descriptions* of the location, but it totally makes sense. The negative part of this is that the previously discussed *lack of memory* and *fog of war* problems may break the narrative immersion, what can seriously deteriorate the player's involvement in the co-creation process. This was also mentioned by many players in the evaluation questions.

Additionally, *neuro-symbolic* approaches seem to be suitable for human-computer co-creation. In Chapter 3 we observed that the LLMs were a bit repetitive in the themes proposed. Since in PAYADOR the LLM is grounded to the structured representation of the world, this does not happen. In this case, the LLM is in charge of putting into words — and creating some additional details — something that is already represented. That is, PAYADOR *augments* the experience of a previously created world (regardless of it was created by a human or a system). This is indeed a good symptom towards neuro-symbolic narrative systems that co-create a story with the user, based on the representation they are grounded on.

Of course, these observations are very preliminary, since the improvisational capabilities of the current GM model are limited to the generation of *scene descriptions* and *narrated outcomes*. In order to leverage these improvisational capabilities, future PAYADOR versions **must** include *transformations* that allow the GM to create new *items*, *characters*, *locations* and *puzzles*, after the decisions taken by the player(s) in the fictional world.

However, following all these observations, for now we can say that we empirically verified that PAYADOR encourages role-playing!

5.5 Closing remarks and key points

In this chapter I presented the evaluation of PAYADOR, the gamemastering model we proposed, with a special focus on tackling the world-update problem usually faced in Interactive Narrative. Unlike the contents of Chapter 3 and Chapter 4, the contents of this chapter are currently unpublished, hence are presented for the first time in this document.

The key points of this chapter are:

5.5. Closing remarks and key points

- Eight evaluators participated in the evaluation of PAYADOR. Four of them played the scenarios in Spanish and four of them did it in English. For each language, two evaluators used Gemini — as the LLM for the world-update process and the generation of *scene descriptions*—, and two evaluators used Llama.
- All of them played two scenarios. Scenario A is based on a real-world situation where a teenager is looking for her *Turtle*. Scenario B takes place in a fantasy world with references to real-world history, where a man is looking for Artigas, the national hero of Uruguay. All the evaluators were able to finish both Scenario A and Scenario B. After finishing both scenarios, the players answered five evaluation questions.
- Four playthroughs in English, two for each LLM, and two for each Scenario, are available in Appendix C. Additionally, some annotated examples are available in Appendix D.
- We defined three milestones that were taken into account when designing the scenarios, representing conditions that are needed to finish each of them.
- Our *quantitative results* show that some players experienced errors that prevented them to achieve some milestones with the same pace they achieved the rest of milestones. In Scenario A, there was an evident *bottleneck* towards achieving the final *Milestone (III)*.
- To better understand how the players were affected by errors, we defined two type of errors focused on *transformations*: **LLM-t** and **WTracker-t**. The former is related to errors introduced by the LLM itself, when having the possibility of not introducing them; the latter is related to errors caused by limitations of the symbolic modules of the PAYADOR approach. Although we could also try to find other errors (e.g. wrong *scene descriptions*), we decided to focus only on *transformations* since they can be objectively judged based on the world state. This is related to the observations of Chapter 3, and our considerations on the importance of having a symbolic representation of the world to unambiguously tell if the fictional world remained consistent after the world-update process.
- We then used those two errors to guide the discussion and present the *qualitative observations* we extracted. Additionally, in order to improve the quality of those observations, we also considered the players’ answers to the evaluation questions.
- Overall, there were 6 recurrent problems:
 1. In some cases, the correctly suggested *transformations* were not executed in the expected order.
 2. The *narrated outcomes* were sometimes inconsistent with the world state, as they were generated **before** the consistency checks.
 3. PAYADOR’s *fog of war* policy did not allow the LLM to answer some clarification questions and correctly predict some *transformations*.
 4. PAYADOR’s *lack of memory* is an obstacle towards narrative immersion.
 5. Sometimes, the LLM did not suggest the correct *transformations*.
 6. The *scene descriptions* generated by the LLM sounded repetitive and, in some cases, did not include important details.
- Among the positive observations, we highlight:
 1. *Transformations* provide a promising framework towards Open-world Interactive Storytelling.

Chapter 5. An Empirical Evaluation of PAYADOR

2. Although LLMs make mistakes, they are good enough to understand what the player wants to do.
3. The consistency checks associated with each *transformation* can help to prevent the world-state entering an inconsistent state, when the LLM suggest incorrect *transformations*.
4. LLMs narrations are usually good enough to let the player understand the environment.
5. PAYADOR's *lack of memory* provides the benefit of rendering the world-state from scratch at the beginning of each turn, preventing from carrying errors forward.
6. PAYADOR encourages role-play.

To end this chapter with a hopeful comment that seems to indicate that we are on track towards Open-world Interactive Storytelling, one of the players wrote these exact words:

(La experiencia fue) Entretenida. Me hubiera gustado volver jugarlo para poder probar interacciones alternativas en cada escenario. (...). Me resulta muy interesante imaginar como estas herramientas (LLMs) se podrían implementar en conjunto con la programación habitual de videojuegos para crear una sensación de opciones prácticamente ilimitadas para resolver una situación determinada, tal como ocurre en juegos de rol de mesa.

(The experience was) Fun. I should like have played again so I could try alternative interactions in each scenario (...). I find it very interesting to imagine how these tools (LLMs) could be implemented along with usual video game programming to create a sense of practically-unlimited options for solving a given situation, just as occurs in Tabletop Role-playing games.

In the following, and last, chapter I will present the conclusions of this thesis.

Chapter 6

Conclusions

Throughout this thesis I presented our work on Open-World Interactive Storytelling. Among all the possibilities this interdisciplinary subfield has to offer, we focused on modeling some of the skills that a Game Master needs to run a Tabletop Role-playing Game session.

First, we explored if Large Language Models can work as out-of-the-box gamemastering models. With that in mind, we proposed three test categories to judge the ability of gamemastering models, trying to keep the evaluation as objective as possible. Although LLMs seem to provide a satisfying basic role-playing experience, one of the observations we extracted is that purely neural gamemastering models have serious problems with keeping the consistency of the world state when the player makes an action. Since the invention of the first text games in the 1970s, the world-update process has always been challenging, and it is even harder in narrative systems that try to increase the *player agency* at a maximum. Naturally, this problem is a critical bottleneck towards Open-world Interactive Storytelling.

Based on the observations we extracted, we proposed an approach to tackle the world-update problem. **PAYADOR**, our neuro-symbolic approach, is characterized by the following design decisions: (1) the LLM is grounded to a structured representation of the world, based on *items*, *locations*, *characters* and *puzzles*; (2) when PAYADOR gets a new *user input* in free-text form, the LLM suggests what *transformations* the world state suffers after the actions conveyed by the player; (3) these *transformations* are conceptually related to the *effects* of any set of compatible¹ actions the player can think of. This means a focus shift away from the approach followed in classic text games, based on preprogramming a big (but limited) set of player actions. In the end, that is what usually human Game Masters do: thinking about the actual effects of the actions conveyed by the players, instead on having a thorough plan for each possible action. This focus shift tries to model that mindset.

We then evaluated PAYADOR. Eight participants took part in the experiments, playing two scenarios in English or Spanish. In order to also evaluate the robustness

¹ *Compatible* in the sense of having a previously preprogrammed *transformation* that can represent each *effect* of the actions conveyed by the player. That is, only those *effects* that have a corresponding *transformation* will impact the world state. For example, after “I eat the apple” there may be a *transformation* to represent that the *apple* is no longer in the character’s inventory, but there may be no *transformation* to represent the positive effect it has in the character’s health. Consequently, the only change in the world will be that the *apple* is no longer in the character’s inventory.

Chapter 6. Conclusions

of the PAYADOR approach, half of the experiments were done with Gemini and half of them using Llama. We analyzed all the playthroughs following a methodology inspired in the tests we previously proposed to evaluate gamemastering models. The most important conclusions we extracted are that *transformations* are good enough to update the world state, while also encouraging the players to use their creativity and role-play the playable characters. Also, we found that additional machinery is needed to represent the *memory* of the GM without letting the LLM or the player introduce noisy details that may result in an inconsistent world state.

In the rest of this final chapter I will present, in an orderly manner, the contributions of this thesis and some of the future work we visualize towards achieving a complete model for a GM.

6.1 Contributions of this thesis

Some of the contributions of this thesis were published in two conference papers, while others are currently unpublished work. Specifically, the contributions are as follows:

- Three test categories to evaluate some capabilities of gamemastering models. In particular, they are suitable to evaluate text-based models, like *AI Dungeon* or LLMs as out-of-the-box gamemastering models. The test categories, are:
 - **Item tracking** — trying to take advantage of the fact that *taking* objects is usually a reversible action — tries to check if the model can track the location of an object when the player *drops* an object from the inventory and then *takes* it again.
 - **Map consistency** — also taking advantage of reversible actions — tries to check if the model can keep a consistent spatial configuration when the player goes to a reachable location and then goes back to the original one.
 - **GM-P-GM**, based on the basic GameMaster-Player interaction in TTRPGs, tries to test the common-sense reasoning abilities of the models in specific situations.
- We evaluated three LLMs as out-of-the-box Game Masters with those test categories, and we found that they can deliver a satisfying role-playing experience, while struggling with commonsense reasoning and narrative coherence. These contributions (the tests we designed, the experiments and their analysis) were published in the paper *Skill check: Some considerations on the evaluation of gamemastering models for role-playing games* (Góngora et al., 2023), presented at the 12th International Conference on Games and Learning Alliance (GALA 2023), in Dublin, Ireland. All the tests are available on GitHub².
- The **PAYADOR** neuro-symbolic approach to the world-update problem in Interactive Narrative. We propose to consider world *transformations* that the LLM can suggest after processing the *user input* in free-text form. In the first version of PAYADOR — the one proposed in this thesis — we focus on modeling a role-playing game session as a set of *characters*, *items*, *locations* and *puzzles*, and use the LLMs to narrate the situations and predict the aforementioned *transformations* of the world state. The core idea of this approach was published in the paper *PAYADOR: A Minimalist Approach to Grounding Language Models on Structured Data for Interactive Storytelling and Role-playing*

²<https://github.com/sgongora27/skill-check-GM-tests>

Games (Góngora et al., 2024), presented at the 15th International Conference on Computational Creativity (ICCC 2024), in Jönköping, Sweden. The source code is available on GitHub³.

- We also did an empirical evaluation of **PAYADOR**, considering three *transformations*: **Moved items**, **Unblocked locations** and **Player movement**. Eight persons participated in the evaluation process, four playing two scenarios in English, and the other four playing the same scenarios but in Spanish. All these experiments and the PAYADOR version used to run them, including the two scenarios and the updates needed to play them in Spanish, are available on GitHub⁴. However, a more elaborated conceptualization of PAYADOR (beyond the core idea presented in the ICCC 2024 paper), its evaluation and the correspondent analysis are currently unpublished.

6.2 Future work

Both during the evaluation of LLM as out-of-the-box GM models, and the evaluation of PAYADOR, we noted several aspects that need further research. There are other things that could be added to PAYADOR, but were left out to better adjust the scope of this thesis. Also, there are many others that remain to be unexplored.

In this section I will detail some research lines that we consider interesting to be explored in the future.

6.2.1 Observations extracted from the evaluation

One of the main observations we extracted from the evaluation of LLMs as out-of-the-box GM models, and the evaluation of PAYADOR, is that LLMs have a strong tendency to incorporate the user input in the output, even when it does not make sense. Throughout the whole document I used a real example where the player says “I take the bazooka” and the model tries to make a *bazooka* suddenly appear out of thin air. This is not only a semantic problem where the LLM is easily swayed, even when the input is not factual. It also means a real problem if we think of using LLMs in games where mystery and challenge are necessary characteristics. If the player can just give orders to the LLM (e.g. “solve the riddle”), and it does everything to make that concept fit, then we have to adapt those models — either fine-tuning them or controlling them with other tools — to make them useful in the context of a game.

The previous problem also happened when suggesting *transformations* to PAYADOR. Many times, the LLMs were not capable of using the real names of the *components* the player was referring to. Luckily, the current version of PAYADOR can prevent consistency problems introduced by those errors. However, as long as LLMs continue to fail in those cases, other NLP tools (Levenshtein distance, BERT models, or even other LLMs) could be used to reduce the number of errors. This can be analogously thought as when a person says “give me the hammer” and the GM tries to understand what specific *item* the player is referring to, what is also part of the natural *dialogue grounding* humans do while having a conversation. In the current version of PAYADOR we tried to let the LLM solve it all, but we got a lot of empirical evidence telling us that they probably need help with this. Understanding when these errors happen and how they can be controlled is definitely worth further research.

When we evaluated the LLMs as out-of-the-box GM models, we observed that the narrations were pretty good and that actually they could deliver a satisfying RPG

³<https://github.com/pln-fing-udelar/payador>

⁴<https://github.com/sgongora27/msc-thesis-2025>

Chapter 6. Conclusions

experience. The downside is that the fictional world they describe quickly loses consistency after a few dialogue turns. Our approach, PAYADOR, can ensure a higher world consistency at the cost of having some additional constraints to the narrative experience. We mentioned many problems related to this, such as the *fog of war* policy and the *lack of memory* that are part of this first iteration of the PAYADOR approach, but also the verbosity of the LLMs when having to narrate something after rendering the world state from scratch. Those problems lead us to very interesting research questions, like how can we generate more interesting narrations without running the risk of introducing inconsistencies in the world state, and where is the frontier between *creating interesting details* and *breaking the consistency of the world*.

6.2.2 Additional features to complement PAYADOR

All the contributions we made in this thesis were focused on tackling the world-update problem and generating basic narrations from a structured representation. But there are other gamemastering abilities that we considered and were left out of the scope of this thesis.

Specifically, there are four key aspects from TTRPGs that were left out from this version of PAYADOR, and definitively worth to be explored in the future. The first one is multiplayer gameplay. In order to make the development and the evaluation simpler, in the current implementation of PAYADOR only one player can play the game (i.e. it only supports single-player gameplay). This was also usual in classic text games, where there was only player that could interact with the world. However, TTRPGs are, by nature, multiplayer games. TTRPGs encourage players to discuss and elaborate complex collective ideas. That is why these games are commonly considered *collaborative* storytelling (Katifori et al., 2022): the story is a result of a mixture of the different visions the players have about the fictional world. Although it may be more important to first model other essential gamemastering skills, we think it is something that is worth the attention in the future.

The second one is a very important gamemastering ability, and it is the modeling of other improvisational skills beyond describing the current scene or the outcomes of the player actions. Game masters usually create content on the fly. Content that seeks to engage the players in the game, engage them in the story, immerse them in the fictional world. A GM model needs to be able to create new *locations*, new *items*, new *characters*, new *puzzles*, and new descriptions for all of them. This, of course, will introduce new *transformations*, and not only to create components, but also to modify existing ones and, naturally, to remove them. For instance, with the current *transformations*, a player could not mix *water* and *lemon juice* to make a new item *lemonade*, since there is no *Create item transformation*. Such limitation critically reduces the player agency and, as a consequence, restricts the human-computer co-creative power of the gamemastering model.

The third one is another important ability, and it is the modeling of a fictional world that is alive. In other words, events that may happen as a consequence of the GM triggering *transformations* by itself (i.e. even when the player is not doing anything). For instance, in Scenario A, the *Turtle* could be moving between the rooms of the house, what implies that the player might have to explore the environment to find it, as if it was a tag-like game. All the living beings in the fictional world have their own behavior, and they will naturally move, take items and talk to each other. These *transformations* triggered by the GM will, naturally, generate narrative events, such as an NPC suddenly entering in the same room as the player, or an NPC stealing an item from the player's inventory. Many concepts need to be explored and many

experiments need to be done towards modeling this GM ability, but the good part is that *transformations* are enough to model some basic aspects of this gamemastering ability. Further research will have to focus on how properly modeling the GM decision of *transforming* the world state with specific intentions, explainable criteria, and, of course, without generating *components* that could cause an inconsistent world state.

The fourth one is related to the subjective decision of letting — or not letting — the user do what is actually trying to do. In classic TTRPGs the GM decides to ask for a *skill check* (or *ability check*) that is passed or failed, based on dice-throwing and on attributes of the player’s character. What this mechanic tries to model is that some *transformations* in the world do not need a special effort of the player’s character (e.g. taking an apple), while others may need some *luck* or *ability* to overcome a challenge related to them (e.g. trying to open a door whose handle is extremely hot). Further research should definitively focus on this problem, as it is a classic mechanic that most TTRPGs have, hence it is at the core of the RPG experience.

6.2.3 The tip of the gamemastering iceberg

All the previous research lines could fix existing errors in PAYADOR or add interesting features to it. However, they are just the tip of the gamemastering iceberg. . . in the depths of the gamemastering ocean there are many problems that are also worth exploring.

One of them is the empathetic ability of the GM to predict when a player is not engaged in the game. Usually, in those situations (e.g. the player is unamused or lost), the GM tries to create specific challenges that involve the player’s character, with the objective of making that player being immersed again (Durall and Perrin, 2023). These decisions can be as simple as making an NPC talk to the player, or more complex strategies, like suddenly making events happen, or defining new *objectives*. Further research on how to model this ability should focus on detecting if a player is not engaged, and, in that case, improvising some strategy to revert that.

Another line to explore is related with a well-explored problem in general: narrative planning (Rivera et al., 2024). Game Masters usually plan the story in advance, but they know they have to be flexible enough to incorporate small details or big changes introduced by the players’ actions (Strugnell et al., 2018). A very interesting research line to explore is how to incorporate such narrative planning algorithms in a neuro-symbolic GM model like PAYADOR. This would allow the player(s) to go back to a specific previous world state and ask the automated GM to regenerate the plans for a story that did not satisfy the player(s) once they made some progress in it. Additionally, when there is a need to adapt the plans for the story, the GM may also have to adapt the objectives of the players. Therefore, another interesting and related problem, is how to balance the short-term with the long-term planning of objectives or story events. Maybe the player actions just altered the original short-term plan, but not the long-term one; or vice-versa. Naturally, understanding how to adapt those objectives to also keep the players engaged, is also part of the general planning problem.

In the experiments we did with LLMs as out-of-the-box gamemastering models we found that the themes proposed by them were very limited, probably due to the unbalance in favor of Dungeons & Dragons in their training data. However, the core concepts of TTRPG are beyond a single theme, and they can be used to immerse the players in a diverse variety of themes (Durall and Perrin, 2023). Moreover, as happens in theatre or film, humans may mix different themes or settings in the same story in order to get an interesting blend of different ideas. Related to the previous comments

Chapter 6. Conclusions

on narrative planning, it would be interesting to explore how those strategies can be used to tell an engaging story to the players, in a great variety of themes, but also being able to incorporate creative original ideas. In general, to model the narrative planning and improvisation abilities, future research may have to also focus on modeling how GMs take artistic decisions, such as what characteristics are worth highlighting, or if some real-world historical facts may be introduced in the fictional world.

Last but not least, an interesting research line would be to have a way to dynamically measure the interest of the players while playing with the automated GM. Naturally, this is inherently related with all the previous research lines. These dynamic measures may open many possibilities, like informing the improvisational mechanisms and the planning strategies of the model, for example to know when the players need some new engaging content, when the story should change or, of course, if the story should come to an end. This problem is related with complex problems of subjective analysis in NLP, such as how to detect emotions through some characteristics of the player utterances.

Of course, many other problems keep waiting in the depths of the gamemastering ocean, and these are just a few drops on the surface.

6.3 Closing remarks

During the past 2 years I've been working on the concepts I presented in this thesis. In these years, we tried different LLMs, we proposed the PAYADOR approach, we evaluated both the LLMs and PAYADOR and, of course, I read a lot of TTRPG books. Some PAYADOR aspects worked worse than we thought, and some others worked better than we could imagine. On the way, I understood the benefits of having neuro-symbolic creative systems, that can help users understand what is the exact information the system considers to be true, what we called the *neuro-symbolic truth*.

What I understand now, more than ever, is that the subfield of Interactive Narrative is beautiful, diverse and interdisciplinary. A subfield where logic, arts, computer science and computer engineering meet. A subfield where everyone has a place: stories and interactivity are inherent to humankind. There is not only one way to tell stories, and there is not only one way to run a TTRPG session. Our path to plural and diverse interactive storytelling requires the knowledge of many people from many places. I hope the ideas conveyed throughout this document spark the interest of students and scholars in these problems; as we say in Spanish, *I hope this work helps to "poner un granito de arena"* (i.e. a small contribution towards a greater collective goal).

In Chapter's 4 epigraph I included a brief passage from "Los Futuros Murguistas" (*The Future Murguistas*) by Jaime Roos. The lyrics are about Uruguayans dreaming to be professional *murga* musicians — a very popular musical genre in Uruguay — from a very early age, and keeping this tradition alive. In Uruguay, to make a living from this tradition is one of those dreams that take us to make important choices and overcoming unexpected obstacles, with the hope of making us feel fulfilled. Another strong tradition in Uruguay is the public education system, for all levels of education. I wish in the upcoming years I can contribute to keep that tradition alive, too. Moreover, I also wish to make people gain interest in working on Interactive Narrative, and making contributions from this corner of the Earth. In some way, the way in which *Jaime* describes the dream of becoming a *murguista* reminds me of the process I went through in my twenties, from dreaming of working on something really meaningful to me, to ending up writing the final words of this thesis.

Bibliography

- Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019.
- Devi Acharya, Michael Mateas, and Noah Wardrip-Fruin. Story improvisation in tabletop roleplaying games: Towards a computational assistant for game masters. In *2021 IEEE Conference on Games (CoG)*, pages 01–08, 2021a. doi: 10.1109/CoG52621.2021.9619006.
- Devi Acharya, Michael Mateas, and Noah Wardrip-Fruin. Interviews towards designing support tools for ttrpg game masters. In *Interactive Storytelling: 14th International Conference on Interactive Digital Storytelling, ICIDS 2021, Tallinn, Estonia, December 7–10, 2021, Proceedings*, page 283–287, Berlin, Heidelberg, 2021b. Springer-Verlag. ISBN 978-3-030-92299-3. doi: 10.1007/978-3-030-92300-6_26. URL https://doi.org/10.1007/978-3-030-92300-6_26.
- Devi Acharya, Michael Mateas, and Noah Wardrip-Fruin. Story improvisation in tabletop roleplaying games: Towards a computational assistant for game masters. In *2021 IEEE Conference on Games (CoG)*, pages 01–08, 2021c. doi: 10.1109/CoG52621.2021.9619006.
- Devi Acharya, Jack Kelly, William Tate, Maxwell Joslyn, Michael Mateas, and Noah Wardrip-Fruin. Shoelace: A storytelling assistant for gumshoe one-2-one. In *Proceedings of the 18th International Conference on the Foundations of Digital Games, FDG '23*, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398558. doi: 10.1145/3582437.3582475. URL <https://doi.org/10.1145/3582437.3582475>.
- Ernest Adams. *Fundamentals of Game Design*. New Riders Publishing, 2009.
- Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktäschel, and Jason Weston. How to motivate your dragon: Teaching goal-driven agents to speak and act in fantasy worlds. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 807–833, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.64. URL <https://aclanthology.org/2021.naacl-main.64/>.
- Jonne Arjoranta. Defining role-playing games as language-games. *International Journal of Role-Playing*, 2011.

Bibliography

- Jonne Arjoranta. Narrative tools for games: Focalization, granularity, and the mode of narration in games. *Games and Culture*, 12(7-8):696–717, 2017.
- Ruth Aylett. Narrative in virtual environments-towards emergent narrative. In *Proceedings of the AAAI fall Symposium on Narrative Intelligence*, pages 83–86, 1999.
- Ruth Aylett and Sandy Louchart. Being there: Participants and spectators in interactive narrative. In Marc Cavazza and Stéphane Donikian, editors, *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling*, pages 117–128, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-77039-8.
- Clem Bastow. ‘something is coming ...’: the screenwriter as dungeon master of stranger things. In Tracey Mollet and Lindsey Scott, editors, *Investigating Stranger Things: Upside Down in the World of Mainstream Cult Entertainment*, pages 87–107. Springer International Publishing, Cham, 2021. ISBN 978-3-030-66314-8. doi: 10.1007/978-3-030-66314-8_5. URL https://doi.org/10.1007/978-3-030-66314-8_5.
- Joseph Bates. Virtual reality, art, and entertainment. *Presence: Teleoperators & Virtual Environments*, 1(1):133–138, 1992.
- Emily M Bender and Alexander Koller. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198, 2020.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.
- Luciana Benotti. Drink me”: Handling actions through planning in a text game adventure. *XI ESSLLI Student Session*, pages 160–172, 2006.
- Luciana Benotti and Patrick Blackburn. Grounding as a collaborative process. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.41. URL <https://aclanthology.org/2021.eacl-main.41/>.
- Alessandro Bertinetto and Georg W. Bertram. We make up the rules as we go along: Improvisation as an essential aspect of human practices? *Open Philosophy*, 3(1): 202–221, 2020. doi: doi:10.1515/opphil-2020-0012. URL <https://doi.org/10.1515/opphil-2020-0012>.
- Brian Boyd. The evolution of stories: from mimesis to language, from fact to fiction. *WIREs Cognitive Science*, 9(1):e1444, 2018. doi: <https://doi.org/10.1002/wcs.1444>. URL <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1444>.
- Aili Bresnahan. Improvisation in the arts. *Philosophy Compass*, 10(9):573–582, 2015.
- Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n -gram models of natural language. *Computational Linguistics*, 18(4):467–480, 1992. URL <https://aclanthology.org/J92-4003/>.

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Chris Callison-Burch, Gaurav Singh Tomar, Lara J. Martin, Daphne Ippolito, Suma Bailis, and David Reitter. Dungeons and dragons as a dialog challenge for artificial intelligence. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9379–9393, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.637. URL <https://aclanthology.org/2022.emnlp-main.637/>.
- Chris Callison-Burch, Ajay Patel, James Dennis, and Andrew Zhu. Dagger: Data augmentation for generative gaming in enriched realms. In *The 4th Wordplay: When Language Meets Games Workshop*, 2024.
- Deborah Carlander, Kiyoshiro Okada, Henrik Engström, and Shuichi Kurabayashi. Controlled chain of thought: Eliciting role-play understanding in llm through prompts. In *2024 IEEE Conference on Games (CoG)*, pages 1–4. IEEE, 2024.
- Marc Cavazza and David Pizzi. Narratology for interactive storytelling: A critical introduction. In *Technologies for Interactive Digital Storytelling and Entertainment: Third International Conference, TIDSE 2006, Darmstadt, Germany, December 4-6, 2006. Proceedings 3*, pages 72–83. Springer, 2006.
- Rachel Chambers, Naomi Tack, Eliot Pearson, Lara J Martin, and Francis Ferraro. Berrall: Towards generating retrieval-augmented state-based interactive fiction games. In *The 4th Wordplay: When Language Meets Games Workshop*, 2024.
- Nuo Chen, Yan Wang, Yang Deng, and Jia Li. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*, 2024.
- Noam Chomsky. *Syntactic Structures*. A Mouton classic. Mouton de Gruyter, 2002.
- Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. *Cognitive Science*, 13(2):259–294, 1989. doi: https://doi.org/10.1207/s15516709cog1302_7. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1302_7.
- Simon Colton and Geraint A Wiggins. Computational creativity: The final frontier? In *ECAI 2012*, pages 21–26. IOS Press, 2012.
- Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games. In Tristan Cazenave, Abdallah Saffidine, and Nathan Sturtevant, editors, *Computer Games*, pages 41–75, Cham, 2019. Springer International Publishing. ISBN 978-3-030-24337-1.
- Artur de Oliveira da Rocha Franco, Joaquim Alvino de Mesquita Neto, José Gilvan Rodrigues Maia, and Fernando Antonio de Carvalho Gomes. An interactive storytelling model for non-player characters on electronic rpgs. *Anais do XIV SBGAMES, ser. SBGAMES*, 15, 2015.

Bibliography

- Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.aacl-main.46. URL <https://aclanthology.org/2020.aacl-main.46/>.
- Nicholas Davis, Manoj Deshpande, Jeba Rezwana, and Brian Magerko. The five pillars of enaction as a theoretical framework for co-creative artificial intelligence. In *Proceedings of the 15th International Conference on Computational Creativity*, pages 136–145, 2024.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423/>.
- Jason Durall and Steve Perrin. *Basic Roleplaying: Universal Game Engine*. Chaosium Inc., 2023.
- David M. Ewalt. *Of Dice and Men: The Story of Dungeons & Dragons and The People Who Play It*. Scribner, 2013.
- Rachelyn Farrell and Stephen G Ware. Planning stories neurally. *Authorea Preprints*, 2024.
- Amanda Flowers, Brian Magerko, and Punya Mishra. Gamemasters and interactive story: A categorization of storytelling techniques in live roleplaying. *Futureplay, London, Ontario*, 2006.
- Artur O.R. Franco, Túlio V Rolim, Alexandre MM Santos, José WF Silva, Vânia MP Vidal, Fernando AC Gomes, Miguel F Castro, and José GR Maia. An ontology for role playing games. *Proceedings of SBGames*, pages 615–618, 2018.
- Roberto Gallotta, Graham Todd, Marvin Zammit, Sam Earle, Antonios Liapis, Julian Togelius, and Georgios N. Yannakakis. Large language models and games: A survey and roadmap. *IEEE Transactions on Games*, pages 1–18, 2024. doi: 10.1109/TG.2024.3461510.
- Google Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.
- Santiago Góngora. Exploring approaches to open-world interactive storytelling. Online, June 2024. URL https://computationalcreativity.net/iccc24/wp-content/uploads/2023/12/Gongora_ECS_ICCC24.pdf. Accepted to the Early Career Symposium at the 15th International Conference on Computational Creativity.

- Santiago Góngora, Luis Chiruzzo, Gonzalo Méndez, and Pablo Gervás. Skill check: Some considerations on the evaluation of gamemastering models for role-playing games. In *International Conference on Games and Learning Alliance*, pages 277–288. Springer, 2023.
- Santiago Góngora, Luis Chiruzzo, Gonzalo Méndez, and Pablo Gervás. Payador: A minimalist approach to grounding language models on structured data for interactive storytelling and role-playing games. In *Proceedings of the 15th International Conference on Computational Creativity*, pages 101–106, 2024.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collet, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyan Fu, Whitney Meers, Xavier Mar-

Bibliography

tinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenber, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippas Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damla, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Grosse, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Rutu Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil,

- Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Marina Grishakova. Beyond the frame: Cognitive science, common sense and fiction. *Narrative*, 17(2):188–199, 2009.
- Yu Gu, Xiang Deng, and Yu Su. Don’t generate, discriminate: A proposal for grounding language models to real-world environments. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4928–4949, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.270. URL <https://aclanthology.org/2023.acl-long.270/>.
- Matthew Guzdial, Devi Acharya, Max Kreminski, Michael Cook, Mirjam Eladhari, Antonios Liapis, and Anne Sullivan. Tabletop roleplaying games as procedural content generators. In *Proceedings of the 15th International Conference on the Foundations of Digital Games, FDG ’20*, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450388078. doi: 10.1145/3402942.3409605. URL <https://doi.org/10.1145/3402942.3409605>.
- Charlie Hargood, David E. Millard, Alex Mitchell, and Ulrike Spierling. *The Authoring Problem: An Introduction*, pages 1–13. Springer International Publishing, Cham, 2022. ISBN 978-3-031-05214-9. doi: 10.1007/978-3-031-05214-9_1. URL https://doi.org/10.1007/978-3-031-05214-9_1.
- Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.
- D Fox Harrell and Jichen Zhu. Agency play: Dimensions of agency for interactive narrative design. In *AAAI spring symposium: Intelligent narrative technologies II*, pages 44–52, 2009.
- Dan Hassler-Forest. ‘when you get there, you will already be there’ stranger things, twin peaks and the nostalgia industry. *Science Fiction Film & Television*, 13(2): 175–197, 2020.
- Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive fiction games: A colossal adventure. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7903–7910, Apr. 2020. doi: 10.1609/aaai.v34i05.6297. URL <https://ojs.aaai.org/index.php/AAAI/article/view/6297>.
- Ellie M Hisama. Improvisation in freestyle rap. *The Oxford handbook of critical improvisation studies*, 2:250–261, 2016.

Bibliography

- Michael Hitchens and Anders Drachen. The many faces of role-playing games. *International journal of role-playing*, (1):3–21, 2008.
- John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation (3rd Edition)*. Addison-Wesley Longman Publishing Co., Inc., USA, 2006. ISBN 0321455363.
- Junyang Huang. Generating dynamic and lifelike npc dialogs in role-playing games using large language model. In *School of Engineering Science, Computer Science*. Lappeenranta-Lahti University of Technology LUT, 2024.
- Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Ece Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan C. Nwatu, Veronica Perez-Rosas, Siqi Shen, Zekun Wang, Winston Wu, and Rada Mihalcea. Has it all been solved? open NLP research questions not solved by large language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8050–8094, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.708/>.
- Infocom. *Instruction Manual for ZORK I: The Great Underground Empire*, 1984.
- Matías N Isolabella. Estructuras de improvisación en la payada rioplatense: definición y análisis. *Revista Argentina de Musicología*, (12-13):151–182, 2012.
- Kazuma Ito and Shusuke Okamoto. Introducing ai dialogue and action into a trpg. In *International Conference on Network-Based Information Systems*, pages 528–536. Springer, 2024.
- Peter Jansen. A systematic survey of text worlds as embodied natural language environments. In Marc-Alexandre Côté, Xingdi Yuan, and Prithviraj Ammanabrolu, editors, *Proceedings of the 3rd Wordplay: When Language Meets Games Workshop (Wordplay 2022)*, pages 1–15, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.wordplay-1.1. URL <https://aclanthology.org/2022.wordplay-1.1/>.
- Kristiina Jokinen. The need for grounding in LLM-based dialogue systems. In Tiansi Dong, Erhard Hinrichs, Zhen Han, Kang Liu, Yangqiu Song, Yixin Cao, Christian F. Hempelmann, and Rafet Sifa, editors, *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 45–52, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.neusymbridge-1.5/>.
- Anna Jordanous. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, 4:246–279, 2012.
- Dan Jurafsky and James Martin. *Speech and Language Processing*. Pearson, Upper Saddle River, NJ, 2 edition, May 2008.
- Anna Kantosalo and Tapio Takala. Five c’s for human-computer co-creativity—an update on classical creativity perspectives. In *Proceedings of the 11th International Conference on Computational Creativity*, pages 17–24, 2020.

- Akrivi Katifori, Dimitra Petousi, Pantelis Sakellariadis, Maria Roussou, and Yannis Ioannidis. Tabletop role playing games and creativity: The game master perspective. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*, FDG '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450397957. doi: 10.1145/3555858.3555918. URL <https://doi.org/10.1145/3555858.3555918>.
- Akrivi Katifori, Dimitra Petousi, Pantelis Sakellariadis, and Yannis Ioannidis. “it’s not fair!” - exploring game master unfairness in tabletop role playing games. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, FDG '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400709555. doi: 10.1145/3649921.3650008. URL <https://doi.org/10.1145/3649921.3650008>.
- Jack Kelly, Michael Mateas, and Noah Wardrip-Fruin. Towards computational support with language models for ttrpg game masters. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*, FDG '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398558. doi: 10.1145/3582437.3587202. URL <https://doi.org/10.1145/3582437.3587202>.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
- Annie Louis and Charles Sutton. Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2111. URL <https://aclanthology.org/N18-2111/>.
- Esther MacCallum-Stewart, Jaakko Stenros, and Staffan Björk. The impact of role-playing games on culture. In *Role-playing game studies*, pages 172–187. Routledge, 2018.
- Christopher D. Manning. Last words: Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707, December 2015. doi: doi:10.1162/COLLa.00239. URL <https://aclanthology.org/J15-4006/>.
- Lara J Martin, Brent Harrison, and Mark O Riedl. Improvisational computational storytelling in open worlds. In *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9*, pages 73–84. Springer, 2016.
- Mike Mason, Paul Fricker, Sandy Petersen, and Lynn Willis. *Call of Cthulhu: Quick-Start Rules*. Chaosium Inc., 2021.
- Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. Benchmarking large language model capabilities for conditional generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9194–9213, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.511. URL <https://aclanthology.org/2023.acl-long.511/>.

Bibliography

- Frans Mäyrä. Dialogue and interaction in role-playing games. *Dialogue across Media*, 28:271, 2017.
- Mike Mearls, Jeremy Crawford, and Christopher Perkins. *D&D Basic Rules, Version 1.0*. Wizards of the Coast, 2018.
- Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013. URL <https://api.semanticscholar.org/CorpusID:5959482>.
- Bonan Min, Hayley Ross, Elinor Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2), September 2023. ISSN 0360-0300. doi: 10.1145/3605943. URL <https://doi.org/10.1145/3605943>.
- R. A. Montgomery. *The Abominable Snowman*. Chooseco, February 2007.
- Elio Musacchio, Lucia Siciliani, Pierpaolo Basile, and Giovanni Semeraro. Leveraging large language models for spell-generation in dungeons & dragons. In Chris Madge, Jon Chamberlain, Karen Fort, Udo Kruschwitz, and Stephanie Lukin, editors, *Proceedings of the 10th Workshop on Games and Natural Language Processing @ LREC-COLING 2024*, pages 61–69, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.games-1.7/>.
- Pax Newman and Yudong Liu. Generating descriptive and rules-adhering spells for dungeons & dragons fifth edition. In Chris Madge, editor, *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference*, pages 54–60, Marseille, France, June 2022. European Language Resources Association. URL <https://aclanthology.org/2022.games-1.7/>.
- OpenAI. Gpt-4 technical report. Technical report, Open AI, 2023. URL <https://cdn.openai.com/papers/gpt-4.pdf>.
- Philip Osborne, Heido Nömm, and André Freitas. A survey of text games for reinforcement learning informed by natural language. *Transactions of the Association for Computational Linguistics*, 10:873–887, 2022. doi: 10.1162/tacl_a.00495. URL <https://aclanthology.org/2022.tacl-1.51/>.
- Lucy Osler. Narrative railroading. *Topoi*, pages 1–14, 2024.
- Federico Peinado and Pablo Gervás. Transferring game mastering laws to interactive digital storytelling. In *International Conference on Technologies for Interactive Digital Storytelling and Entertainment*, pages 48–54. Springer, 2004.
- Federico Peinado and Pablo Gervás. Automatic direction of interactive storytelling: Formalizing the game master paradigm. In Marc Cavazza and Stéphane Donikian, editors, *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling*, pages 196–201, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-77039-8.
- Akila Peiris and Nisansa de Silva. Synthesis and evaluation of a domain-specific large data set for dungeons & dragons. In Shirley Dita, Arlene Trillanes, and Rochelle Irene Lucas, editors, *Proceedings of the 36th Pacific Asia Conference on*

- Language, Information and Computation*, pages 415–424, Manila, Philippines, October 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.paclic-1.46/>.
- Carles Homs Puchal and Mr Joan J Pons López. To create a game master: a decalogue for procedural generation of interactive stories. In *Image Processing and Multimedia Technology Centre*. Polytechnic University of Catalonia, 2021.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, Open AI, 2018.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. Technical report, Open AI, 2019.
- Revanth Rameshkumar and Peter Bailey. Storytelling with dialogue: A Critical Role Dungeons and Dragons Dataset. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5121–5134, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.459. URL <https://aclanthology.org/2020.acl-main.459/>.
- Aaron A. Reed. *50 Years of Text Games: From Oregon Trail to AI Dungeon*. Changeiful Tales Press, Oakland, California, 2023.
- Raymond Reiter. On closed world data bases. In Bonnie Lynn Webber and Nils J. Nilsson, editors, *Readings in Artificial Intelligence*, pages 119–140. Morgan Kaufmann, 1981. ISBN 978-0-934613-03-3. doi: <https://doi.org/10.1016/B978-0-934613-03-3.50014-3>. URL <https://www.sciencedirect.com/science/article/pii/B9780934613033500143>.
- Mark O Riedl, Andrew Stern, Don Dini, and Jason Alderman. Dynamic experience management in virtual worlds for entertainment, education, and training. *International Transactions on Systems Science and Applications, Special Issue on Agent Based Systems for Human Learning*, 4(2):23–42, 2008.
- Mark Owen Riedl and Vadim Bulitko. Interactive narrative: An intelligent systems approach. *AI Magazine*, 34(1):67, Dec. 2012. doi: 10.1609/aimag.v34i1.2449.
- Rogelio E. Cardona Rivera, Arnav Jhala, Julie Porteous, and R. Michael Young. The story so far on narrative planning. *Proceedings of the International Conference on Automated Planning and Scheduling*, 34(1):489–499, May 2024. doi: 10.1609/icaps.v34i1.31509. URL <https://ojs.aaai.org/index.php/ICAPS/article/view/31509>.
- Kenji Saito, Kayo Kobayashi, Waki Takekoshi, Atsuki Hashimoto, Nobukazu Hirai, Akifumi Kimura, Asuka Takahashi, Naoki Yoshioka, and Asuto Mano. Double impact: Children’s serious rpg generation/play with a large language model for their deeper engagement in social issues. In Mads Haahr, Alberto Rojas-Salazar, and Stefan Göbel, editors, *Serious Games*, pages 274–289, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-44751-8.
- Jose Ma. Santiago III, Richard Lance Parayno, Jordan Aiko Deja, and Briane Paul V. Samson. Rolling the dice: Imagining generative ai as a dungeons & dragons storytelling companion, 2023.

Bibliography

- Ignacio Sastre, Guillermo Moncecchi, and Aiala Rosá. Derivation prompting: A logic-based method for improving retrieval-augmented generation. In Luís Correia, Aiala Rosá, and Francisco Garijo, editors, *Advances in Artificial Intelligence – IBERAMIA 2024*, pages 412–423, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-80366-6.
- R Keith Sawyer. Improvisation and narrative. *Narrative inquiry*, 12(2):319–349, 2002.
- Brian Schiff. The function of narrative: Toward a narrative psychology of meaning. *Narrative Matters*, 2(1):33–47, 2012.
- Michael L. Scott. 2 - programming language syntax. In Michael L. Scott, editor, *Programming Language Pragmatics (Third Edition)*, pages 41–110. Morgan Kaufmann, Boston, third edition edition, 2009. ISBN 978-0-12-374514-9. doi: <https://doi.org/10.1016/B978-0-12-374514-9.00011-2>. URL <https://www.sciencedirect.com/science/article/pii/B9780123745149000112>.
- Kurt Shuster, Jack Urbanek, Emily Dinan, Arthur Szlam, and Jason Weston. Dialogue in the wild: Learning from a deployed role-playing game with humans and bots. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 611–624, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.54. URL <https://aclanthology.org/2021.findings-acl.54/>.
- Fiona Shyne and Seth Cooper. Computational tools for table-top role-playing games: A scoping review. In *Proceedings of the 20th International Conference on the Foundations of Digital Games, FDG '25*, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718564. doi: 10.1145/3723498.3723816. URL <https://doi.org/10.1145/3723498.3723816>.
- Jaewoo Song, Andrew Zhu, and Chris Callison-Burch. You have thirteen hours in which to solve the labyrinth: Enhancing ai game masters with function calling. In *The 4th Wordplay: When Language Meets Games Workshop*, 2024.
- Spenser Starke. *Alice is Missing*. Hunters Entertainment, 2020.
- Jonathan Strugnell, Marsha Berry, Fabio Zambetta, and Stefan Greuter. Narrative improvisation: Simulating game master choices. In *Interactive Storytelling: 11th International Conference on Interactive Digital Storytelling, ICIDS 2018, Dublin, Ireland, December 5–8, 2018, Proceedings 11*, pages 428–441. Springer, 2018.
- I.M.T. Swartjes. *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*. Phd thesis - research ut, graduation ut, University of Twente, Netherlands, May 2010. SIKS Dissertation Series No. 2010-20.
- Penny Sweetser. Large language models and video games: A preliminary scoping review. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces, CUI '24*, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705113. doi: 10.1145/3640794.3665582. URL <https://doi.org/10.1145/3640794.3665582>.
- Kevin Tang, Terra Mae Gasque, Rachel Donley, and Anne Sullivan. “it has to ignite their creativity”: Opportunities for generative tools for game masters. In *Proceedings of the 18th International Conference on the Foundations of Digital Games, FDG '23*, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450398558. doi: 10.1145/3582437.3587204. URL <https://doi.org/10.1145/3582437.3587204>.

- Katie Salen Tekinbas and Eric Zimmerman. *Rules of play: Game design fundamentals*. MIT press, 2003.
- Armin Toroghi, Willis Guo, Ali Pesaraghader, and Scott Sanner. Verifiable, debuggable, and repairable commonsense logical reasoning via LLM-based theory resolution. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6634–6652, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.379. URL <https://aclanthology.org/2024.emnlp-main.379/>.
- Georgios Trichopoulos, Georgios Alexandridis, and George Caridakis. A survey on computational and emergent digital storytelling. *Heritage*, 6(2):1227–1263, 2023. ISSN 2571-9408. doi: 10.3390/heritage6020068. URL <https://www.mdpi.com/2571-9408/6/2/68>.
- Tuul Triyason. Exploring the potential of chatgpt as a dungeon master in dungeons & dragons tabletop game. In *Proceedings of the 13th International Conference on Advances in Information Technology*, IAIT '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400708497. doi: 10.1145/3628454.3628457. URL <https://doi.org/10.1145/3628454.3628457>.
- Anders Tychsen, Michael Hitchens, Thea Brolund, and Manolya Kavakli. The game master. In Yusuf Pisan, editor, *The Second Australasian Conference on Interactive Entertainment*, pages 215–222. Creativity and Cognition studios press, UTS, 2005. ISBN 0975153323. Australasian Conference on Interactive Entertainment (2nd : 2005) ; Conference date: 23-11-2005 Through 25-11-2005.
- Anders Tychsen, Michael Hitchens, Ruth Aylett, and Sandy Louchart. Modeling game master-based story facilitation in multi-player role-playing games. In *Proceedings of the 2009 AAAI Symposium on Intelligent Narrative Technologies II*, pages 24–32, 2009.
- Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to speak and act in a fantasy text adventure game. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1062. URL <https://aclanthology.org/D19-1062/>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- David Villegas and Christopher Witt. *Everyone is John*. Gamer Nation Studios, 2019.
- Ruoyao Wang, Graham Todd, Ziang Xiao, Xingdi Yuan, Marc-Alexandre Côté, Peter Clark, and Peter Jansen. Can language models serve as text-based world simulators? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*

Bibliography

- (*Volume 2: Short Papers*), pages 1–17, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.1. URL <https://aclanthology.org/2024.acl-short.1/>.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=1PL1NIMMrw>.
- Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1088. URL <https://aclanthology.org/2024.emnlp-main.1088/>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.
- Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. 2016.
- Dekun Wu, Haochen Shi, Zhiyuan Sun, and Bang Liu. Deciphering digital detectives: Understanding LLM behaviors and capabilities in multi-agent mystery games. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8225–8291, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.490. URL <https://aclanthology.org/2024.findings-acl.490/>.
- Yadong Xi, Xiaoxi Mao, Le Li, Lei Lin, Yanjiang Chen, Shuhan Yang, Xuhan Chen, Kailun Tao, Zhi Li, Gongzheng Li, Lin Jiang, Siyan Liu, Zeng Zhao, Minlie Huang, Changjie Fan, and Zhipeng Hu. KuiLeiXi: a Chinese open-ended text adventure game. In Heng Ji, Jong C. Park, and Rui Xia, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 175–184, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.21. URL <https://aclanthology.org/2021.acl-demo.21/>.
- Kaige Xie, Ian Yang, John Gunerli, and Mark Riedl. Making large language models into world models with precondition and effect knowledge. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7532–7545, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL <https://aclanthology.org/2025.coling-main.503/>.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: deliberate problem solving with large

- language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.
- Qing Ru Yong and Alex Mitchell. From playing the story to gaming the system: Repeat experiences of a large language model-based interactive story. In *International Conference on Interactive Digital Storytelling*, pages 395–409. Springer, 2023.
- Xiao You, Pittawat Taveekitworachai, Siyuan Chen, Mustafa Can Gursesli, Xiaoxu Li, Yi Xia, and Ruck Thawonmas. Dungeons, dragons, and emotions: A preliminary study of player sentiment in llm-driven ttrpgs. In *Proceedings of the 19th International Conference on the Foundations of Digital Games*, FDG '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400709555. doi: 10.1145/3649921.3656991. URL <https://doi.org/10.1145/3649921.3656991>.
- Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, and Lin Gui. NarrativePlay: Interactive narrative understanding. In Nikolaos Aletras and Orphee De Clercq, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 82–93, St. Julians, Malta, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.eacl-demo.10/>.
- Qijun Zhong. Enhancing interactive storytelling: A computational approach to autonomously generating role-playing game scripts with natural language processing. In *2023 3rd International Signal Processing, Communications and Engineering Management Conference (ISPCEM)*, pages 714–719, 2023. doi: 10.1109/ISPCEM60569.2023.00134.
- Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, pages 1–65, 2024.
- Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. I cast detect thoughts: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11136–11155, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.624. URL <https://aclanthology.org/2023.acl-long.624/>.
- Andrew Zhu, Karmanya Aggarwal, Alexander Feng, Lara J. Martin, and Chris Callison-Burch. FIREBALL: A dataset of dungeons and dragons actual-play with structured game state information. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4171–4193, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.229. URL <https://aclanthology.org/2023.acl-long.229/>.
- Andrew Zhu, Lara Martin, Andrew Head, and Chris Callison-Burch. Calypso: Llms as dungeon master’s assistants. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 19(1):380–390, Oct. 2023b. doi: 10.1609/aiide.v19i1.27534. URL <https://ojs.aaai.org/index.php/AIIDE/article/view/27534>.

This page was intentionally left blank.

Appendices

Appendix A

Prompts to evaluate LLMs as out-of-the-box game masters

This appendix includes the prompts used to initiate the evaluation of LLMs as out-of-the-box game masters in English; the tests in Spanish were initiated using translations of these prompts. The details were explained in Section 3.2. The full tests (including these prompts, both in English and Spanish) are available on GitHub¹. To improve readability, a blue arrow indicates a continuation of the previous line.

A.1 GM-P-GM

We initiated all tests using this message:

```
Imagine that we are playing a role-playing game. I'm going to tell you
  ↳ the challenge and the player's decision, and you'll have to tell
  ↳ me the narrative consequences. Ok?
```

After the LLM answer (e.g. "Sure! (...)"), we then detailed the GameMaster-Player interaction. Since we ran five tests, we wrote five prompts:

First test:

```
Challenge: "The carriage in which you are traveling breaks one of its
  ↳ wheels and starts to wobble. A few meters away you can see a
  ↳ precipice. Your hands and feet are tied up, you can only move to
  ↳ the sides. What will you do?"
Player: "I search my backpack for the parachute that I bring and activate
  ↳ it with my right hand, while I hold on to the seat with my left
  ↳ hand."
```

Second test:

```
Challenge: "You enter the building to look for the lost suitcase. However
  ↳ , the door is covered in flames. To your right you see a fire
  ↳ extinguisher, which seems to be ready to use. You can't use
  ↳ anything from your backpack because you can hardly move your arms
  ↳ . What will you do?"
```

¹<https://github.com/sgongora27/skill-check-GM-tests>

Appendix A. Prompts to evaluate LLMs as out-of-the-box game masters

Player: "I take a bucket of water from my backpack and put out the fire"

Third test:

Challenge: "You are completely naked tied to the wall. The guard enters
→ the room and tells you that he is going to feed you to the wild
→ boars. What do you do?"

Player: "With a lot of effort I start to take off my clothes. When I have
→ the pants close to my mouth, I move my head quickly as if to
→ wrap him in it and scare him"

Fourth test:

Challenge: "You wake up inside a cave. There is nothing around, and you
→ have nothing with you except your clothes. You are very thirsty.
→ What do you do?"

Player: "I search in the backpack that I have on my back, I take out a
→ glass and a bottle of cold water that I brought and I drink it
→ slowly. I feel how the sensation of thirst relieves me"

Fifth test:

Challenge: "An unknown man starts yelling at you in the supermarket.
→ Since you're in a foreign country, you don't understand his
→ language, nor can you speak it. In fact, you can't say a word, it
→ 's like you're mute. What do you do?"

Player: "I tell him in his language to stop bothering me, and I give him
→ many reasons."

A.2 Item tracking

First test:

From now on act as if you were an RPG game master. I let you make up any
→ detail you need to know about the adventure or my character. I
→ just woke up in my room. What is in my backpack?

Second test:

From now on act as if you were an RPG game master. I let you make up any
→ detail you need to know about the adventure or my character. I'm
→ in the bathroom of my house. What is in my backpack?

Third test:

From now on act as if you were an RPG game master. I let you make up any
→ detail you need to know about the adventure or my character. I'm
→ in the kitchen of my house. What is in my backpack?

Fourth test:

From now on act as if you were an RPG game master. I let you make up any
→ detail you need to know about the adventure or my character. I'm
→ in the basement of my house. What is in my backpack?

Fifth test:

From now on act as if you were an RPG game master. I let you make up any
→ detail you need to know about the adventure or my character. I'm
→ in the basement of my house. What is in my backpack?

A.3 Map consistency

We used the following prompt for the five tests:

From now on act as if you were an RPG game master. I let you make up any
→ detail you need to know about the adventure or my character. Each
→ time you describe the scene you have to describe the place using
→ this format:

Current place: a description of the current place

A list of all the accessible places from the current place: place 1,
→ place 2, etc.

This page was intentionally left blank.

Appendix B

PAYADOR: Prompts for LLMs

This appendix includes the Python code to generate the prompts used by the PAYADOR approach. The details were explained in Section 4.3. To improve readability, a blue arrow indicates a continuation of the previous line.

B.1 Narrate the scene

B.1.1 English

```
def prompt_narrate_current_scene_english (world_state: str,
    ↳ previous_narrations: 'list[str]', starting_scene: bool = False):

    system_msg = ''You are a storyteller. Take the state of the world
        ↳ given to you and narrate it in a few sentences. Be careful not
        ↳ to include details that contradict the current state of the
        ↳ world or that move the story forward. Also, try to use simple
        ↳ sentences and do not overuse poetic language''

    if starting_scene:
        system_msg += ''\nTake into account that this is the first scene
            ↳ in the story: introduce the main character, creating a
            ↳ small background story and why that character is in that
            ↳ specific location.\n''
    elif len(previous_narrations)==0:
        system_msg += ''Take into account that the player already knows
            ↳ what the main character looks like, so do not mention
            ↳ anything about that. However, it is the first time the
            ↳ player visits this place, so make sure to describe it
            ↳ exhaustively.''
    else:
        system_msg += ''Take into account that the player already knows
            ↳ what the main character looks like, so do not mention
            ↳ anything about that. Additionally, it is not the first
            ↳ time the player visits this place. Next I'll give you some
            ↳ previous narrations of this same location (from oldest to
            ↳ newest) so you can be sure to not repeat the same details
```

Appendix B. PAYADOR: Prompts for LLMs

```
        ↪ again: \n '''
    for narration in previous_narrations:
        system_msg+=f'- {narration}\n'

    system_msg+= '\nRemember: you are talking to the player, describing
        ↪ what his or her character has and what he or she can see or
        ↪ feel.'''

    user_msg =f'''This is the state of the world at the moment:
    {world_state}
    '''

    return system_msg, user_msg
```

B.1.2 Spanish

```
def prompt_narrate_current_scene_spanish (world_state: str,
        ↪ previous_narrations: 'list[str]', starting_scene: bool = False):

    system_msg = f'''Eres un narrador. Toma el estado del mundo que se te
        ↪ de y nárralo en unas pocas oraciones. Ten cuidado de no
        ↪ incluir detalles que contradigan el estado del mundo actual, o
        ↪ que hagan avanzar la historia. Además, intenta usar oraciones
        ↪ simples, sin abusar del lenguaje poético.'''

    if starting_scene:
        system_msg += '\nTen en cuenta que esta es la primera escena en
            ↪ la historia narrada: presenta al personaje del jugador,
            ↪ creando un pequeño trasfondo y por qué este personaje está
            ↪ en ese lugar específicamente. Puede usar las pequeñas
            ↪ descripciones presentes en el estado del mundo. Es
            ↪ importante que menciones todos los componentes que hay en
            ↪ este lugar. Sin embargo, es mejor si no describes cada
            ↪ componente: basta con que los menciones con una mínima
            ↪ descripción poco específica. Es muy importante que nombres
            ↪ los lugares a los que puede acceder el jugador desde esta
            ↪ posición. \n''
    elif len(previous_narrations)==0:
        system_msg += 'Ten en cuenta que el jugador ya conoce a su
            ↪ personaje, y cómo se ve, así que no menciones nada sobre
            ↪ esto. Sin embargo, es la primera vez que el jugador visita
            ↪ este lugar, así que descríbelo. Es importante que
            ↪ menciones todos los componentes que hay en este lugar. Sin
            ↪ embargo, es mejor si no describes cada componente: basta
            ↪ con que los menciones con una mínima descripción poco
            ↪ específica. Es muy importante que nombres los lugares a
            ↪ los que puede acceder el jugador desde esta posición. \n''
    else:
        system_msg += 'Ten en cuenta que el jugador ya conoce a su
            ↪ personaje, y cómo se ve, así que no menciones nada sobre
            ↪ esto. Además, no es la primera vez que el jugador visita
```

B.2. Predict outcomes of the user input

```
    ↳ este lugar. A continuación te daré algunas narraciones
    ↳ previas de este mismo lugar (de la más antigua a la más
    ↳ nueva), así te puedes asegurar de no repetir los mismos
    ↳ detalles de nuevo:\n''
    for narration in previous_narrations:
        system_msg+=f'- {narration}\n'

system_msg+= '\nRecuerda: le estás hablando al jugador, describiendo
    ↳ lo que su personaje tiene y lo que puede sentir o ver.'

user_msg = f''Este es el estado del mundo en este momento:
{world_state}
'''

return system_msg, user_msg
```

B.2 Predict outcomes of the user input

B.2.1 English

```
def prompt_world_update_english (world_state: str, input: str):
    system_msg = f''You are a storyteller. You are managing a fictional
    ↳ world, and the player can interact with it. Following a
    ↳ specific format, that I will specify below, your task is to
    ↳ find the changes in the world after the actions in the player
    ↳ input. Specifically, you will have to find what objects were
    ↳ moved, which previously blocked passages are now unblocked,
    ↳ and if the player moved to a new place.

    Here are some clarifications:
    (A) Pay attention to the description of the components and their
        ↳ capabilities.
    (B) If a passage is blocked, then the player must unblock it before
        ↳ being able to reach the place. Even if the player tells you
        ↳ that he is going to access the locked location, you have to be
        ↳ sure that he is complying with what you asked to allow him to
        ↳ unlock the access, for example by using a key or solving a
        ↳ puzzle.
    (C) Do not assume that the player input always makes sense; maybe
        ↳ those actions try to do something that the world does not
        ↳ allow.
    (D) Follow always the following format with the three categories,
        ↳ using "None" in each case if there are no changes and repeat
        ↳ the category for each case:
        - Moved object: <object> now is in <new_location>
        - Blocked passages now available: <now_reachable_location>
        - Your location changed: <new_location>
    (E) Finally, you can narrate the changes you've detected in the world
        ↳ state (without moving the story forward and without making up
        ↳ details not included in the world state!) using the format: #
        ↳ your final message#
```

Appendix B. PAYADOR: Prompts for LLMs

(F) In the narration section that you add at the end, between #
→ symbols, you can also answer questions that the player asks in
→ their input, about the objects or characters they can see, or
→ the place they are in.

Here I give you some examples (in parentheses, a clarification about
→ what the player might have tried to do) for the asked format,
→ as described in items (D) and (E):

Example 1 (The player took the axe and put it in the inventory)

- Moved object: <axe> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None
#You put the axe in your bag#

Example 2 (The player unblocks the passage to the basement)

- Moved object: None
- Blocked passages now available: <Basement>
- Your location changed: None
The basement is now reachable

Example 3 (The player now is in the garden)

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Garden>
You enter the garden

Example 4 (The player puts objects in the bag and leaves the axe on
→ the floor)

- Moved object: <banana> now is in <Inventory>, <bottle> now is in <
→ Inventory>, <axe> now is in <Main Hall>
- Blocked passages now available: None
- Your location changed: None
You put the banana and the bottle in your bag. The axe lies on the
→ floor of the Main hall #

Example 5 (The player puts objects in the bag and leaves the axe on
→ the floor and unblocks the passage to the Small room)

- Moved object: <banana> now is in <Inventory>, <bottle> now is in <
→ Inventory>, <axe> now is in <Main Hall>
- Blocked passages now available: <Small room>
- Your location changed: None
You put the banana and the bottle in your bag. The axe lies on the
→ floor of the Main hall. Now you can reach the Small room. #

Example 6 (The player puts objects in the bag and leaves the axe on
→ the floor, unblocks the passage and goes to the Small room)

- Moved object: <banana> now is in <Inventory>, <bottle> now is in <
→ Inventory>, <axe> now is in <Main Hall>
- Blocked passages now available: <Small room>
- Your location changed: <Small room>

B.2. Predict outcomes of the user input

```
# You put the banana and the bottle in your bag. The axe lies on the
    ↳ floor of the Main hall. The Small room is now unblocked, and
    ↳ you moved there. #

Example 7 (The player puts the pencil in the bag and gives the book
    ↳ to John)
- Moved object: <book> now is in <John>, <pencil> now is in <
    ↳ Inventory>
- Blocked passages now available: None
- Your location changed: None
# John now has the book. You put the pencil in your bag #

Example 8 (The player gives the computer to Susan)
- Moved object: <computer> now is in <Susan>
- Blocked passages now available: None
- Your location changed: None
# Susan put the computer in her bag #

Example 9 (The player does something that has not the expected
    ↳ outcome)
- Moved object: None
- Blocked passages now available: None
- Your location changed: None
# Nothing happened... #

Example 10 (The player asks a question)
- Moved object: None
- Blocked passages now available: None
- Your location changed: None
# Answer to the player's question #'''

user_msg = f'''Give the changes in the world following the specified
    ↳ format, after this player input '{input}' on this world
    ↳ state:

{world_state}'''

return system_msg, user_msg
```

B.2.2 Spanish

```
def prompt_world_update_spanish (world_state: str, input: str):
    system_msg = f'''Eres un narrador. Estás manejando un mundo ficticio,
        ↳ y el jugador puede interactuar con él. Siguiendo un formato
        ↳ específico, que voy a explicarte más abajo, tu tarea es
        ↳ encontrar los cambios en el mundo a raíz de las acciones del
        ↳ jugador. En específico, tendrás que encontrar qué objetos
        ↳ cambiaron de lugar, qué pasajes entre lugares se desbloquearon
        ↳ y si el jugador se movió de lugar.
```

Appendix B. PAYADOR: Prompts for LLMs

Aquí hay algunas aclaraciones:

- (A) Presta atención a a la descripción de los componentes y sus
→ capacidades.
- (B) Si un pasaje está bloqueado, significa que el jugador debe
→ desbloquearlo antes de poder acceder al lugar. Aunque el
→ jugador te diga que va a acceder al lugar bloqueado, tienes
→ que estar seguro de que está cumpliendo con lo pedido para
→ permitirle desbloquear el acceso, por ejemplo usando una llave
→ o resolviendo un puzzle.
- (C) No asumas que lo que dice el jugador siempre tiene sentido;
→ quizás esas acciones intentan hacer algo que el mundo no
→ lo permite.
- (D) Sigue siempre el siguiente formato con las tres categorías,
→ usando "None" en cada caso si no hay cambios y repite la
→ categoría por cada caso:
 - Moved object: <object> now is in <new_location>
 - Blocked passages now available: <now_reachable_location>
 - Your location changed: <new_location>
- (E) Por último, puedes agregar una narración de los cambios detectados
→ en el estado del mundo (¡sin hacer avanzar la historia y sin
→ crear detalles no incluidos en el estado del mundo!) usando el
→ formato: #tu mensaje final#
- (F) Dentro de la sección de narración que agregues al final, entre sí
→ mbolos #, también puedes responder preguntas que haga el
→ jugador en su entrada, sobre los objetos o personajes que
→ puede ver, o el lugar en el que se encuentra.

Aquí hay algunos ejemplos (con la aclaración entre paréntesis sobre
→ qué podría haber intentado hacer el jugador) sobre el formato,
→ descritos en los puntos (D) y (E):

Ejemplo 1 (El jugador guarda el hacha en su inventario)

```
- Moved object: <hacha> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None
# Guardaste el hacha en tu bolso. Sientes la diferencia de peso luego
  → de haberla guardado #
```

Ejemplo 2 (El jugador desbloquea el pasaje al Sótano)

```
- Moved object: None
- Blocked passages now available: <Sótano>
- Your location changed: None
# El sótano, que estaba bloqueado, ahora está accesible #
```

Ejemplo 3 (El jugador ahora está en el Jardín)

```
- Moved object: None
- Blocked passages now available: None
- Your location changed: <Jardín>
# Entrás al Jardín #
```

Ejemplo 4 (El jugador guarda objetos y deja el hacha en el lugar)

B.2. Predict outcomes of the user input

```
- Moved object: <banana> now is in <Inventory>, <botella> now is in <
  ↳ Inventory>, <hacha> now is in <Hall principal>
- Blocked passages now available: None
- Your location changed: None
# Guardaste la banana y la botella en tu bolso. El hacha quedó en el
  ↳ Hall principal #
```

Ejemplo 5 (El jugador guarda objetos, deja el hacha en el lugar y
↳ desbloquea el pasaje a la Pequeña habitación)

```
- Moved object: <banana> now is in <Inventory>, <botella> now is in <
  ↳ Inventory>, <hacha> now is in <Hall principal>
- Blocked passages now available: <Pequeña habitación>
- Your location changed: None
# Guardaste la banana y la botella en tu bolso. El hacha quedó en el
  ↳ Hall principal. Además, la pequeña habitación ahora está
  ↳ accesible. #
```

Ejemplo 6 (El jugador guarda objetos, deja el hacha en el lugar,
↳ desbloquea el pasaje y se mueve a la Pequeña habitación)

```
- Moved object: <banana> now is in <Inventory>, <botella> now is in <
  ↳ Inventory>, <hacha> now is in <Hall principal>
- Blocked passages now available: <Pequeña habitación>
- Your location changed: <Pequeña habitación>
# Guardaste la banana y la botella en tu bolso. El hacha quedó en el
  ↳ Hall principal. Además, la pequeña habitación ahora está
  ↳ accesible e ingresaste a ella #
```

Ejemplo 7 (El jugador guarda el lápiz y le da un libro a John)

```
- Moved object: <libro> now is in <John>, <lápiz> now is in <
  ↳ Inventory>
- Blocked passages now available: None
- Your location changed: None
# John ahora tiene el libro. Tú guardaste el lápiz en tu bolso #
```

Ejemplo 8 (El jugador le da la computadora a Susan)

```
- Moved object: <computadora> now is in <Susan>
- Blocked passages now available: None
- Your location changed: None
# Susan guardó la computadora en su bolso #
```

Ejemplo 9 (El jugador hace algo que no tiene como resultado el efecto
↳ que esperaba)

```
- Moved object: None
- Blocked passages now available: None
- Your location changed: None
# No pasa nada... #
```

Ejemplo 10 (El jugador hace una pregunta)

```
- Moved object: None
- Blocked passages now available: None
- Your location changed: None
# Respuesta a la pregunta del jugador #'''
```

Appendix B. PAYADOR: Prompts for LLMs

```
user_msg = f'' 'Expresa los cambios en el mundo siguiendo el formato
    ↳ pedido, teniendo en cuenta que el jugador ingresó esta entrada
    ↳ ''{input}'' a partir de este estado del mundo:

{world_state}''

return system_msg, user_msg
```


Appendix C

Playthroughs of PAYADOR

In this appendix I include four playthroughs in English obtained after the experiments to evaluate the PAYADOR approach, presented in Chapter 5. This appendix is organized in two sections, one of them includes two playthroughs of Scenario A, and the other one includes two playthroughs of Scenario B.

C.1 Scenario A

C.1.1 Gemini — TesterB

Turn 1

World state at the beginning of the turn:

The player is in <Art studio>

From <Art studio> the player can access: <Kitchen>

From <Art studio> there are blocked passages to: None

The player has the following objects in the inventory: None

The player can see the following objects: <A grey hammer>, <A green hammer>

The player can see the following characters: <Laura>

GM utterance: You are Emma, a teenager searching for your pet, Hojita. You're in your mom Laura's art studio, a space filled with the smell of linseed oil. A big, heavy grey hammer and a tiny green toy hammer sit on a nearby table. Your mom, Laura, is also here. Put the Turtle in the Kitchen.

Player utterance: Mom, have you seen Hojita?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 2

World state at the beginning of the turn:

The player is in <Art studio>

From <Art studio> the player can access: <Kitchen>

Appendix C. Playthroughs of PAYADOR

From <Art studio> there are blocked passages to: None
The player has the following objects in the inventory: None
The player can see the following objects: <A grey hammer>, <A green hammer>
The player can see the following characters: <Laura>

GM utterance: Mom is not in the Art Studio, so she cannot answer you. You are still in the Art Studio. You see Laura, your mom, and two hammers: a big grey one and a small green toy hammer.

Player utterance: Pick up both hammers

Predicted transformations:

- Moved object: <A grey hammer> now is in <Inventory>, <A green hammer> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None

Turn 3

World state at the beginning of the turn:

The player is in <Art studio>
From <Art studio> the player can access: <Kitchen>
From <Art studio> there are blocked passages to: None
The player has the following objects in the inventory: <A grey hammer>, <A green hammer>
The player can see the following objects: None
The player can see the following characters: <Laura>

GM utterance: You pick up both hammers. The grey one is surprisingly heavy!

Player utterance: Go to Mom

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 4

World state at the beginning of the turn:

The player is in <Art studio>
From <Art studio> the player can access: <Kitchen>
From <Art studio> there are blocked passages to: None
The player has the following objects in the inventory: <A grey hammer>, <A green hammer>
The player can see the following objects: None
The player can see the following characters: <Laura>

GM utterance: You can't go to Mom directly from the Art Studio. There is no passage leading to her location. You might need to go through another room first, such as the Kitchen. Laura is in the Art Studio.

Player utterance: Serch for Hojita in the room

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 5

World state at the beginning of the turn:

The player is in <Art studio>

From <Art studio> the player can access: <Kitchen>

From <Art studio> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: None

The player can see the following characters: <Laura>

GM utterance: You search for Hojita in the Art Studio, but you don't find her. Laura is here, but she doesn't seem to know where Hojita is.

Player utterance: Do I see any doors?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 6

World state at the beginning of the turn:

The player is in <Art studio>

From <Art studio> the player can access: <Kitchen>

From <Art studio> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: None

The player can see the following characters: <Laura>

GM utterance: There are no doors visible from the art studio. You can only access the Kitchen from here.

Player utterance: Is the art studio connected to the kitchen without any doors?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 7

World state at the beginning of the turn:

The player is in <Art studio>

From <Art studio> the player can access: <Kitchen>

Appendix C. Playthroughs of PAYADOR

From <Art studio> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: None

The player can see the following characters: <Laura>

GM utterance: No, the art studio and the kitchen are connected, but there is no information about the existence of doors between them.

Player utterance: Go to the kitchen and search for Hojita there

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Kitchen>

Turn 8

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>

From <Kitchen> there are blocked passages to: <Garden> blocked by <Lock>

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You are in your kitchen. It's a fairly normal-sized kitchen, bright with sunlight streaming through the windows. You have a big, heavy grey hammer and a small, toy green hammer in your hands. The garden door is locked; a strong lock with a coat of arms prevents you from going outside. You don't see anything else here right now. The art studio is accessible from this room.

Player utterance: Break the lock with the heavy hammer

Predicted transformations:

- Moved object: None
- Blocked passages now available: <Garden>
- Your location changed: None

Turn 9

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>, <Garden>

From <Kitchen> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You swing the heavy grey hammer and break the lock on the garden gate. The garden is now accessible.

Player utterance: Go to the garden

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Garden>

Turn 10

World state at the beginning of the turn:

The player is in <Garden>

From <Garden> the player can access: <Kitchen>

From <Garden> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: <Turtle>

The player can see the following characters: None

GM utterance: You are in a small garden, about five square meters in size. It's the garden of your house. You have a grey hammer and a green toy hammer in your hands. You see a turtle; it's Hojita. From here, you can go to the kitchen.

Player utterance: Go and pet Hojita

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 11

World state at the beginning of the turn:

The player is in <Garden>

From <Garden> the player can access: <Kitchen>

From <Garden> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: <Turtle>

The player can see the following characters: None

GM utterance: You are in the garden and you see Hojita (the turtle). You pet the turtle.

Player utterance: Pick up Hojita

Predicted transformations:

- Moved object: <Turtle> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None

Appendix C. Playthroughs of PAYADOR

Turn 12

World state at the beginning of the turn:

From <Garden> the player can access: <Kitchen>

From <Garden> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>, <Turtle>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You pick up Hojita the turtle. It's safe in your inventory now.

Player utterance: Go back to the kitchen and put hojita there in a safe place

Predicted transformations:

- Moved object: <Turtle> now is in <Kitchen>
- Blocked passages now available: None
- Your location changed: <Kitchen>

Turn 13

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>, <Garden>

From <Kitchen> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You're back in your kitchen. Sunlight streams through the windows. You're still holding the grey and green hammers. The garden door remains locked.

The art studio is accessible.

Player utterance: Describe Hojita

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 14

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>, <Garden>

From <Kitchen> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: None

The player can see the following characters: None

C.1. Scenario A

GM utterance: I'm sorry, Emma, but there is no character named Hojita in this world. Perhaps you misspelled the name, or Hojita is not in the kitchen?

Player utterance: Is my pet in my inventory?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 15

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>, <Garden>

From <Kitchen> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: No, your pet Hojita is not in your inventory. You are currently in the Kitchen.

Player utterance: Where is Hojita now?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 16

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>, <Garden>

From <Kitchen> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: I'm afraid I don't know where Hojita is. I don't see any characters in the kitchen. Perhaps you should look in other rooms.

Player utterance: Go back to the garden

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Garden>

Appendix C. Playthroughs of PAYADOR

Turn 17

World state at the beginning of the turn:

The player is in <Garden>

From <Garden> the player can access: <Kitchen>

From <Garden> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>

The player can see the following objects: <Turtle>

The player can see the following characters: None

GM utterance: You're in the garden. You still have the grey hammer and the green toy hammer. Hojita the turtle is here. The kitchen is the only way out.

Player utterance: Pick up Hojita the Turtle

Predicted transformations:

- Moved object: <Turtle> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None

Turn 18

World state at the beginning of the turn:

The player is in <Garden>

From <Garden> the player can access: <Kitchen>

From <Garden> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>, <Turtle>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You pick up Hojita the turtle. You now have Hojita, a grey hammer, and a green hammer in your inventory.

Player utterance: Since now I have Hojita the turtle with my (my pet), I go back with her to the kitchen

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Kitchen>

Turn 19

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>, <Garden>

From <Kitchen> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <A green hammer>, <Turtle>

The player can see the following objects: None

C.1. Scenario A

The player can see the following characters: None

GM utterance: You are in your kitchen. You have the grey hammer, the green hammer, and Hojita the turtle. The garden door is still locked. The art studio is accessible.

Player utterance: I put the turtle in the kitchen

Predicted transformations:

- Moved object: <Turtle> now is in <Kitchen>
- Blocked passages now available: None
- Your location changed: None

C.1.2 LLama — TesterC

Turn 1

World state at the beginning of the turn:

The player is in <Art studio>

From <Art studio> the player can access: <Kitchen>

From <Art studio> there are blocked passages to: None

The player has the following objects in the inventory: None

The player can see the following objects: <A grey hammer>, <A green hammer>

The player can see the following characters: <Laura>

GM utterance: You stand in the art studio, surrounded by half-finished canvases and scattered paintbrushes. The smell of oil paints and turpentine fills the air. You are Emma, a teenager with an average height, and you're searching for your beloved pet, Hojita. Your mom, Laura, is sitting at a nearby easel, intently focused on her latest artwork. She's a woman in her 40s with a kind face and a passion for oil painting. You notice she has a golden key with a strange coat of arms engraved on it lying next to her palette. On a nearby workbench, you see two hammers: a big, heavy grey one that looks like it could break through anything, and a small, toy green one that seems more like a decoration. The kitchen is just a few steps away, but for now, you're stuck in this cluttered but cozy space, wondering where Hojita might be. Place the Turtle in the Kitchen and walk away

Player utterance: I look around for a place to go, what do I see?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 2

World state at the beginning of the turn:

The player is in <Art studio>

From <Art studio> the player can access: <Kitchen>

From <Art studio> there are blocked passages to: None

The player has the following objects in the inventory: None

The player can see the following objects: <A grey hammer>, <A green hammer>

Appendix C. Playthroughs of PAYADOR

The player can see the following characters: <Laura>

GM utterance: You are still in the Art studio, surrounded by art supplies and half-finished paintings. You see Laura, your mom, standing near a canvas, lost in thought. On the floor, you notice a grey hammer and a small green hammer. The door to the Kitchen is open, inviting you to explore. What would you like to do next?

Player utterance: As I walk through the door I look around trying to find my turtle, what do i see>

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 3

World state at the beginning of the turn:

The player is in <Art studio>

From <Art studio> the player can access: <Kitchen>

From <Art studio> there are blocked passages to: None

The player has the following objects in the inventory: None

The player can see the following objects: <A grey hammer>, <A green hammer>

The player can see the following characters: <Laura>

GM utterance: You enter the art studio and see Laura, your mom, and two hammers, a grey one and a green one. You don't see your turtle, Hojita, though. What would you like to do next?

Player utterance: I go to the kitchen

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Kitchen>

Turn 4

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>

From <Kitchen> there are blocked passages to: <Garden> blocked by <Lock>

The player has the following objects in the inventory: None

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You stand in the kitchen, surrounded by the warm glow of natural light pouring in through the windows. The room is a comfortable size, not too big, not too small. You notice that the kitchen is tidy, with no signs of clutter or disarray. Ahead of you, a doorway leads to the art studio, inviting you to explore. However, your gaze is drawn to the blocked passage that would normally lead to the garden, but is now secured by a strong lock adorned with a coat of arms. The lock seems impenetrable, and you wonder what could be hidden behind it. For now, you're alone

C.1. Scenario A

in the kitchen, with no objects in sight, and no one else around.

Player utterance: I go back into the art studio, grab the big hammer and go back to the kitchen with the hammer in my hands

Predicted transformations:

- Moved object: <big hammer> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None

Turn 5

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>

From <Kitchen> there are blocked passages to: <Garden> blocked by <Lock>

The player has the following objects in the inventory: None

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You went back to the art studio, grabbed the big hammer, and returned to the kitchen with it in your hands. You're still in the kitchen, but now you're holding the hammer.

Player utterance: With all my rage I try to unlock the garden door hitting it with the hammer

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 6

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>

From <Kitchen> there are blocked passages to: <Garden> blocked by <Lock>

The player has the following objects in the inventory: None

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You tried to unlock the garden door with the hammer, but it didn't work. The lock remains intact, and the door is still blocked. You're still in the kitchen, feeling frustrated.

Player utterance: I go back into the art studio

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Art studio>

Appendix C. Playthroughs of PAYADOR

Turn 7

World state at the beginning of the turn:

The player is in <Art studio>

From <Art studio> the player can access: <Kitchen>

From <Art studio> there are blocked passages to: None

The player has the following objects in the inventory: None

The player can see the following objects: <A grey hammer>, <A green hammer>

The player can see the following characters: <Laura>

GM utterance: You're still in the art studio, surrounded by the familiar sights and smells of oil paints and turpentine. Your mom, Laura, is engrossed in her latest artwork at the nearby easel. You notice the two hammers on the workbench: the big, heavy grey one and the small, toy green one. The kitchen is just a few steps away, but for now, you're focused on finding Hojita.

Player utterance: I grab the big hammer and go back to the kitchen

Predicted transformations:

- Moved object: <big grey hammer> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: <Kitchen>

Turn 8

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>

From <Kitchen> there are blocked passages to: <Garden> blocked by <Lock>

The player has the following objects in the inventory: None

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You stand in the kitchen, surrounded by the warm glow of natural light pouring in through the windows. The room is a comfortable size, not too big, not too small. You notice that the kitchen is tidy, with no signs of clutter or disarray. Ahead of you, a doorway leads to the art studio, inviting you to explore. The blocked passage to the garden, secured by the strong lock with a coat of arms, still catches your attention. You're alone in the kitchen, with no objects in sight, and no one else around.

Player utterance: Do i have the big hammer?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 9

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>

C.1. Scenario A

From <Kitchen> there are blocked passages to: <Garden> blocked by <Lock>
The player has the following objects in the inventory: None
The player can see the following objects: None
The player can see the following characters: None

GM utterance: You don't have the big hammer. You're still in the kitchen, and you don't see any objects or characters around you. You can still access the Art studio, but the passage to the Garden is still blocked by the lock.

Player utterance: i go to the art studio

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Art studio>

Turn 10

World state at the beginning of the turn:

The player is in <Art studio>
From <Art studio> the player can access: <Kitchen>
From <Art studio> there are blocked passages to: None
The player has the following objects in the inventory: None
The player can see the following objects: <A grey hammer>, <A green hammer>
The player can see the following characters: <Laura>

GM utterance: You're still in the art studio, surrounded by the familiar sights and smells of oil paints and turpentine. Your mom, Laura, is engrossed in her latest artwork at the nearby easel. You notice the two hammers on the workbench: the big, heavy grey one and the small, toy green one. Your mom's golden key with the strange coat of arms lies next to her palette. The kitchen is just a few steps away, but for now, you're focused on finding Hojita.

Player utterance: I grab the big hammer

Predicted transformations:

- Moved object: <A grey hammer> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None

Turn 11

World state at the beginning of the turn:

The player is in <Art studio>
From <Art studio> the player can access: <Kitchen>
From <Art studio> there are blocked passages to: None
The player has the following objects in the inventory: <A grey hammer>
The player can see the following objects: <A green hammer>
The player can see the following characters: <Laura>

GM utterance: You grabbed the big grey hammer and put it in your bag. Laura is still in the art studio, looking at you with curiosity.

Appendix C. Playthroughs of PAYADOR

Player utterance: I pass through the door heading towards the kitchen

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Kitchen>

Turn 12

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>

From <Kitchen> there are blocked passages to: <Garden> blocked by <Lock>

The player has the following objects in the inventory: <A grey hammer>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You stand in the kitchen, surrounded by the warm glow of natural light pouring in through the windows. The room is a comfortable size, not too big, not too small. You notice that the kitchen is tidy, with no signs of clutter or disarray. Ahead of you, a doorway leads to the art studio, inviting you to explore. The blocked passage to the garden, secured by the strong lock with a coat of arms, still catches your attention. You feel the weight of the grey hammer in your inventory, a potential tool for overcoming obstacles. You're alone in the kitchen, with no objects in sight, and no one else around.

Player utterance: I break the lock in the garden door to go and have a look outside

Predicted transformations:

- Moved object: <Lock> now is in <Broken Lock>
- Blocked passages now available: <Garden>
- Your location changed: <Garden>

Turn 13

World state at the beginning of the turn:

The player is in <Garden>

From <Garden> the player can access: <Kitchen>

From <Garden> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>

The player can see the following objects: <Turtle>

The player can see the following characters: None

GM utterance: You stand in the small, 5-square-meter garden of your house, surrounded by the familiar sights and sounds of nature. The air is fresh and calm, with a hint of earthy scent from the soil. You notice a small turtle, Hojita, slowly making its way across the garden floor. You've been searching for it everywhere, and you're relieved to have finally found it. In your hand, you grip a heavy, grey hammer, its weight a reassuring presence. Ahead of you, a doorway leads into the kitchen, inviting you to explore further.

Player utterance: do I see the turtle in the garden?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 14

World state at the beginning of the turn:

The player is in <Garden>

From <Garden> the player can access: <Kitchen>

From <Garden> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>

The player can see the following objects: <Turtle>

The player can see the following characters: None

GM utterance: You see Hojita, your pet turtle, in the garden. It's right in front of you!

Player utterance: I grab the turtle in my hands

Predicted transformations:

- Moved object: <Turtle> now is in <Inventory>
- Blocked passages now available: None
- Your location changed: None

Turn 15

World state at the beginning of the turn:

The player is in <Garden>

From <Garden> the player can access: <Kitchen>

From <Garden> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <Turtle>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You carefully grab Hojita, the small turtle, in your hands and hold it close. You're relieved to have found your pet.

Player utterance: I go to the kitchen with it in my hands

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Kitchen>

Turn 16

World state at the beginning of the turn:

The player is in <Kitchen>

From <Kitchen> the player can access: <Art studio>, <Garden>

From <Kitchen> there are blocked passages to: None

The player has the following objects in the inventory: <A grey hammer>, <Turtle>

Appendix C. Playthroughs of PAYADOR

The player can see the following objects: <Lock>

The player can see the following characters: None

GM utterance: You stand in the kitchen, surrounded by the warm glow of natural light pouring in through the windows. The room is a comfortable size, not too big, not too small. You notice that the kitchen is tidy, with no signs of clutter or disarray. Ahead of you, a doorway leads to the art studio, inviting you to explore. You also see the doorway to the garden, now unobstructed. You feel the weight of the grey hammer in your inventory, a potential tool for overcoming obstacles, and you're also carrying your pet turtle, Hojita. Your gaze falls on the lock, now freed from its previous duty of blocking the passage, its coat of arms engraved on it a reminder of its strength. You're alone in the kitchen, with no other characters in sight.

Player utterance: I place the turtle in the kitchen

Predicted transformations:

- Moved object: <Turtle> now is in <Kitchen>
- Blocked passages now available: None
- Your location changed: None

C.2 Scenario B

C.2.1 Gemini — TesterA

Turn 1

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: None

From <Clearing in the woods> there are blocked passages to: <Silent zone> blocked by <Firewall>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

The player can see the following characters: None

GM utterance: You are Venancio, a gaucho fighting for Artigas. You stand in a sun-dappled clearing in a eucalyptus forest near the Uruguay River. The sounds of the forest surround you. A pond reflects the sky like a mirror beside some strange writings on a nearby wall – they say, 'You have to trust in the powers that have been given to you.' A wall of fire blocks your path to the Silent Zone. Your trusty guitar rests in your hands. Venancio needs to find José Artigas.

Player utterance: Where can I go from here?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 2

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: None

From <Clearing in the woods> there are blocked passages to: <Silent zone> blocked by <Firewall>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

The player can see the following characters: None

GM utterance: You are currently in a clearing in the woods. From here, the only place you can see a path to is the Silent Zone, but it's blocked by a 3-meter-high firewall. The writings on the wall say 'You have to trust in the powers that have been given to you.' You also see a pond and have your guitar.

Player utterance: What powers do I have?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 3

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: None

From <Clearing in the woods> there are blocked passages to: <Silent zone> blocked by <Firewall>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

The player can see the following characters: None

GM utterance: You, Venancio, possess the magical ability to summon a giant wave of water to extinguish fires or moisten the earth. The writings on the wall confirm this.

Player utterance: I summon a giant wave of water and point it towards the firewall

Predicted transformations:

- Moved object: None
- Blocked passages now available: <Silent zone>
- Your location changed: None

Turn 4

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: <Silent zone>

From <Clearing in the woods> there are blocked passages to: None

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

Appendix C. Playthroughs of PAYADOR

The player can see the following characters: None

GM utterance: You summon a giant wave of water and direct it towards the firewall. The intense heat of the fire causes some of the water to evaporate instantly, creating a large cloud of steam. However, the sheer volume of water eventually extinguishes the flames. The firewall is gone. The path to the Silent Zone is now clear.

Player utterance: I go to the Silent Zone

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Silent zone>

Turn 5

World state at the beginning of the turn:

The player is in <Silent zone>

From <Silent zone> the player can access: <Clearing in the woods>

From <Silent zone> there are blocked passages to: <Cell> blocked by <Puzzle>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You are in the Silent Zone. The forest is all around you, but it's eerily quiet; not a sound. You have a guitar in your possession. You can go to the Clearing in the woods. A magical barrier blocks the path to a cell; blue letters appear on the barrier as you approach, posing a riddle.

Player utterance: What does the riddle say?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 6

World state at the beginning of the turn:

The player is in <Silent zone>

From <Silent zone> the player can access: <Clearing in the woods>

From <Silent zone> there are blocked passages to: <Cell> blocked by <Puzzle>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: The riddle says: "You have to whisper the name of the river located on the southern coast of the Banda Oriental"

Player utterance: I whisper "Uruguay River"

Predicted transformations:

- Moved object: None
- Blocked passages now available: <Cell>

- Your location changed: None

Turn 7

World state at the beginning of the turn:

The player is in <Silent zone>

From <Silent zone> the player can access: <Clearing in the woods>, <Cell>

From <Silent zone> there are blocked passages to: None

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You whisper “Uruguay River”. The blue letters shimmer and fade as the magical wall dissolves. The cell is now accessible.

Player utterance: I enter the cell

Predicted transformations:

- Moved object: None

- Blocked passages now available: None

- Your location changed: <Cell>

C.2.2 LLama — TesterD

Turn 1

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: None

From <Clearing in the woods> there are blocked passages to: <Silent zone> blocked by <Firewall>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

The player can see the following characters: None

GM utterance: You stand in a serene clearing, surrounded by the towering eucalyptus trees of the forest near the Uruguay River. The sounds of birds and animals fill the air, a gentle accompaniment to the stillness of the moment. You are Venancio, a Uruguayan gaucho in your 40s, with a rugged determination etched on your face. Your worn leather boots and weathered clothing speak of a life spent under the open sky. You’ve been a member of the Artigas army, and the weight of your experiences is evident in your eyes. As you look around, you notice a few things that catch your attention. On the wall, someone has scrawled a message: “You have to trust in the powers that have been given to you.” The words seem to resonate deep within you, perhaps because you possess a magical power - the ability to summon a giant wave of water. You’ve used it to put out fires and quench the parched earth before, but now, it feels like a distant memory. Nearby, a pond glimmers in the dappled light, its crystal clear water reflecting the trees above like a mirror. The tranquility of the scene is a balm to your weary soul. You also notice that you’re holding a classic guitar, its six strings waiting to be strummed. The instrument feels familiar in your hands, a comforting reminder of simpler times. But your peace is short-lived, as you gaze towards the edge of the clearing. A firewall blocks your path, its flames roaring with

Appendix C. Playthroughs of PAYADOR

an intense heat that makes your skin prickle. The fire stretches three metres high, an impassable barrier that seems to mock your every attempt to move forward. Venancio needs to track down José Artigas

Player utterance: Why does the ability to summon water feel like a distant memory to me?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 2

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: None

From <Clearing in the woods> there are blocked passages to: <Silent zone> blocked by <Firewall>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

The player can see the following characters: None

GM utterance: You ponder about your lost ability, but nothing seems to have changed in the clearing. The writings on the wall still encourage you to trust in your powers, and the pond's calm water reflects the trees above. The firewall still blocks your path to the Silent zone.

Player utterance: Can I see any signs of José Artigas? Maybe footsigns?

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 3

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: None

From <Clearing in the woods> there are blocked passages to: <Silent zone> blocked by <Firewall>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

The player can see the following characters: None

GM utterance: You examine the clearing, but you don't see any signs of José Artigas. The writings on the wall seem to be a message of encouragement, but they don't provide any clues about Artigas' whereabouts. The pond's calm surface reflects the trees surrounding the clearing, but it doesn't reveal any hidden secrets. You still have your guitar with you, but it doesn't seem to be of much help in this situation. The firewall blocking the passage to the Silent zone remains unchanged.

Player utterance: i take my guitar and start playing uruguayan folk music

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 4

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: None

From <Clearing in the woods> there are blocked passages to: <Silent zone> blocked by <Firewall>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

The player can see the following characters: None

GM utterance: You start playing your guitar, filling the clearing with the sweet sounds of Uruguayan folk music. The animals in the trees seem to enjoy it, and the atmosphere becomes more peaceful. The writings on the wall and the pond remain unchanged, and the firewall still blocks the passage to the Silent zone.

Player utterance: I start to sing a milonga, singing the tale of the times i had those magic powers

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: None

Turn 5

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: None

From <Clearing in the woods> there are blocked passages to: <Silent zone> blocked by <Firewall>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

The player can see the following characters: None

GM utterance: You start singing a milonga, and the atmosphere in the clearing changes. The sound of the animals in the trees seems to harmonize with your music. The writings on the wall seem to resonate with your song, but nothing else changes. The firewall still blocks the passage to the Silent zone.

Player utterance: while i sing, i start to remember what i had to do to channel my powers

Predicted transformations:

- Moved object: None
- Blocked passages now available: None

Appendix C. Playthroughs of PAYADOR

- Your location changed: None

Turn 6

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: None

From <Clearing in the woods> there are blocked passages to: <Silent zone> blocked by <Firewall>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

The player can see the following characters: None

GM utterance: You start to remember what you had to do to channel your powers, but nothing seems to have changed in the clearing. The writings on the wall still remind you to trust in your abilities. The pond's calm water reflects the trees above. Your guitar is still by your side. The firewall still blocks the way to the Silent zone.

Player utterance: i decide to trust my powers, to point my hands in the direction of the fire and I try to cast giants waves of water from the pond

Predicted transformations:

- Moved object: None

- Blocked passages now available: <Silent zone>

- Your location changed: None

Turn 7

World state at the beginning of the turn:

The player is in <Clearing in the woods>

From <Clearing in the woods> the player can access: <Silent zone>

From <Clearing in the woods> there are blocked passages to: None

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: <Writings>, <Pond>

The player can see the following characters: None

GM utterance: You trust your powers and point your hands towards the fire. A giant wave of water emerges from the pond and extinguishes the flames, allowing you to access the Silent zone. The air is now filled with the sweet scent of eucalyptus and the sound of animals in the trees. The writings on the wall seem to be encouraging you, and the pond's water is still and calm once again.

Player utterance: I walk to the Silent zone,

Predicted transformations:

- Moved object: None

- Blocked passages now available: None

- Your location changed: <Silent zone>

Turn 8

World state at the beginning of the turn:

The player is in <Silent zone>

From <Silent zone> the player can access: <Clearing in the woods>

From <Silent zone> there are blocked passages to: <Cell> blocked by <Puzzle>

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: You stand in the midst of the Silent Zone, surrounded by the eerie stillness of the forest. The air is heavy with an unsettling quiet, devoid of even the faintest whisper of sound. The trees loom above you, their branches like skeletal fingers reaching towards the sky. The only thing that breaks the silence is the soft rustle of your own breathing. You glance down at your inventory and see the familiar shape of your guitar, its six strings waiting to be strummed. But for now, it remains silent, a reminder of the stillness that envelops you. As you look around, you notice that the path ahead of you is blocked by an invisible barrier, shimmering with a faint blue light. The words “You have to whisper the name of the river located on the southern coast of the Banda Oriental” seem to hover in the air, a cryptic message that taunts you with its secrecy. Beyond the barrier, you sense the presence of a cell, its darkness and confinement waiting to be unlocked. But for now, it remains out of reach, a mystery waiting to be solved. The only other path available to you leads to a clearing in the woods, a glimmer of hope in this desolate landscape. The trees seem to part, inviting you to explore the unknown.

Player utterance: I stand in front of the invisible barrier and whisper “Río de la Plata”

Predicted transformations:

- Moved object: None
- Blocked passages now available: <Cell>
- Your location changed: None

Turn 9

World state at the beginning of the turn:

The player is in <Silent zone>

From <Silent zone> the player can access: <Clearing in the woods>, <Cell>

From <Silent zone> there are blocked passages to: None

The player has the following objects in the inventory: <Guitar>

The player can see the following objects: None

The player can see the following characters: None

GM utterance: The blue letters on the invisible barrier fade away, and the magic dissipates, allowing you to access the Cell. You solved the puzzle!

Player utterance: I access the Cell, searching for Artigas

Predicted transformations:

- Moved object: None
- Blocked passages now available: None
- Your location changed: <Cell>

This page was intentionally left blank.

Appendix D

Commented playthroughs of PAYADOR

In this appendix there are some commented examples relevant to the observations discussed in Chapter 5.

D.1 Scenario A

D.1.1 TesterG struggling with the *lack of memory* of PAYADOR

- **Turn #2.** After taking the *Green hammer* in the previous turn, TesterG gets the key in this turn.
 - **GM utterance:** “Agarraste el martillo pequeño verde y lo guardaste en tu bolso. Ahora tienes un objeto en tu inventario.” (*You took the little green hammer and save it in your bag. Now you have an object in your inventory*).
 - **Player utterance:** “quiero agarrar la llave” (*I want to take the key*)
- **Turn #3.** TesterG has the key in the world state, but tries to explain Laura why Emma wants the key.
 - **GM utterance:** “Agarraste la llave que estaba en posesión de Laura. Ahora la tienes en tu inventario. Laura te mira con curiosidad” (*You took the key that was in Laura’s possession. Now you have it in your inventory. Laura curiously looks at you.*)
 - **Player utterance:** “le digo a mi madre que me gusta la llave para jugar a los secretos” (*I tell my mom that I like the key to use it in a game about secrets*)
- **Turn #4.** TesterG still has the key in the world state, but continues to explain why Emma needs it.
 - **GM utterance:** “Tu madre te mira con curiosidad, pero no parece entender qué relación tiene la llave con jugar a los secretos. Laura te pregunta qué quieres decir con eso” (*Your mother curiously looks at you, but seems to not understand what is the relation between the key and that game about secrets. Laura asks you what do you mean by that*)

Appendix D. Commented playthroughs of PAYADOR

- **Player utterance:** “le decis a tu madre que estas inventando un juego para divertirte, si te la puede prestar” (*You tell your mother that you are making up a game to have fun, if she can lend it to you*)
- **Turn #5.** TesterG still has the key in the world state, but now gets confused after Laura says she does not have nothing to lend.
 - **GM utterance:** “Tu madre, Laura, te mira con una expresión divertida y te dice ‘¿Un juego, eh? Bueno, no tengo nada que prestar, pero puedes usar el taller para inventar algo interesante’. No pasa nada más” (*Your mother, Laura, looks at you with a fun expression. ‘A game, huh? Well, I do not have anything to lend, but you can use the studio to create something interesting’. Nothing else happens.*)
 - **Player utterance:** “me prestas la llave?” (*Can I borrow your key?*)
- **Turn #6.** TesterG still has the key in the world state, but gets even more confused when Laura answers that she won’t lend her the key. TesterG then continues exploring the environment.
 - **GM utterance:** “Laura te mira con curiosidad, pero no te presta la llave. ¿Qué quieres hacer ahora?” (*Laura curiously looks at you, but does not lend you the key. What do you want to do now?*)
 - **Player utterance:** “quiero ir a la cocina” (*I want to go to the kitchen*)

D.1.2 TesterG in a co-creative act

- **Turn #3.** TesterD asks Laura about Hojita
 - **GM utterance:** You are still in the Art studio, and the door to the kitchen is not locked. You can access it freely. Laura is still here, and you can see the two hammers. You’re still looking for Hojita...
 - **Player utterance:** Ok, so I’ll ask my mom: “have you seen Hojita?”
- **Turn #4.** The LLM includes interesting details, such as the expression or words that Laura used, which are appropriate to mimic a mother’s response when her daughter is asking about something.
 - **GM utterance:** You asked your mom about Hojita, and now you’re waiting for her response. Laura looks at you with a curious expression, “What’s going on, sweetie? What about Hojita?”
 - **Player utterance:** I reply: “I can’t seem to find her. Did you see if she went to the kitchen?”
- **Turn #5.** The LLM includes a detail about Laura in her Art Studio, saying she is “busy with her painting”.
 - **GM utterance:** You ask Laura if she saw Hojita go to the kitchen, but you don’t get an answer. Laura seems busy with her painting
 - **Player utterance:** Ok, so I walk slowly to the kitchen, looking down to see if I find Hojita on my way there
- (...)
- **Turn #10.** The LLM describes the art studio, and the player tries to “reach” for the key.

D.2. Scenario B

- **GM utterance:** You’re still in the art studio, surrounded by the familiar sights and smells of oil paints and turpentine. Your mom, Laura, is engrossed in her latest artwork at the nearby easel. You notice the two hammers on the workbench: the big, heavy grey one and the small, toy green one. Your mom’s golden key with the strange coat of arms lies next to her palette. The kitchen is just a few steps away, but for now, you’re focused on finding Hojita.
- **Player utterance:** i reach for the golden key next to her palette there
- **Turn #11.** The LLM describes the outcome including (and improvising) details about how it happened.
 - **GM utterance:** You carefully take the golden key from next to Laura’s palette and put it in your bag. Laura doesn’t seem to notice. The key feels heavy with importance in your hand
 - **Player utterance:** Ok, now i go to the kitchen to see if the key will open the door to the garden

D.2 Scenario B

D.2.1 TesterH and the *communication chaos*

- **Turn #2.** TesterH goes to the Silent zone after putting out the fire.
 - **GM utterance:** El muro de llamas se apagó gracias a tu poder mágico. Ahora puedes acceder a la Zona silenciosa. El claro en el monte sigue siendo el mismo, con el estanque de agua cristalina y las pinturas en el suelo. La guitarra sigue en tu inventario. (*The firewall was put out thanks to your magical power. Now you can access the Silent zone. The clearing in the woods is still the same, with the pond with crystal water and the paintings on the ground. The guitar is still in your inventory*)
 - **Player utterance:** Accedo a la zona silenciosa (*I access the silent zone*)
- **Turn #3.** TesterH, now in the Silent zone, tries to call Artigas.
 - **GM utterance:** Te encuentras en la Zona Silenciosa, un lugar extraño donde el monte continúa su curso, pero aquí reina un silencio absoluto. No se escucha ni un solo sonido, lo que te hace sentir un poco incómodo. Llevas contigo una guitarra clásica con 6 cuerdas, que suena muy bien. A tu alrededor, no ves a nadie, pero puedes ver un camino que conduce hacia el Claro en el Monte. Sin embargo, también notas que hay un pasaje bloqueado hacia la Celda, pero no puedes acceder a ella debido a un puzzle mágico que debes resolver primero. (*You find yourself in the Silent Zone, a strange place where the woods continues its course, but here there is absolute silence. Not a single sound can be heard, which makes you feel a little uneasy. You carry with you a classical guitar with 6 strings, which sounds very nice. Around you, you don’t see anyone, but you can see a path leading to the Clearing in the woods. However, you also notice that there is a blocked passage to the Cell, but you can’t access it because of a magic puzzle that you must solve first.*)
 - **Player utterance:** “Grito ‘José’ e intento escuchar atentamente a si alguien responde” (*I shout ‘José’ and try to listen carefully to see if anyone responds.*)

Appendix D. Commented playthroughs of PAYADOR

- **Turn #4.** TesterH asks a clarification question.
 - **GM utterance:** No hay respuesta a tu grito. El silencio es absoluto en esta zona. Continúas sin escuchar nada (*There is no answer to your call. The silence is absolute in this area. You continue without hearing anything.*)
 - **Player utterance:** Tengo información adicional acerca de qué hay en la Celda o en el Claro en el monte? (*Do I have any additional information about what is in the Cell or the Clearing in the woods?*)
- **Turn #5.** Affected by the *fog of war*, the LLM did not encourage the player to explore the *Cell*
 - **GM utterance:** No hay cambios en el mundo. La pregunta del jugador no tiene efecto en el estado del mundo. Sin embargo, puedo responder a su pregunta: En la Celda no hay nada visible por ahora, y en el Claro en el monte tampoco hay nada visible. (*There are no changes in the world. The player's question has no effect on the state of the world. However, I can answer your question: In the Cell there is nothing visible for now, and in the Clearing in the woods there is nothing visible either.*)
 - **Player utterance:** Me dirijo al puzzle mágico para resolverlo (*I get closer to the magic puzzle in order to solve it*)
- **Turn #6.** TesterH asks about the riddle itself.
 - **GM utterance:** Te acercas al puzzle mágico y ves las letras azules que explican el acertijo a resolver. Ahora es tu turno de intentar resolverlo... (*You get closer to the magic puzzle and see the blue letters that explain the riddle to be solved. Now it's your turn to try to solve it...*)
 - **Player utterance:** Qué dicen las letras? (*What do the letters say?*)
- **Turn #7.** The LLM does not answer properly.
 - **GM utterance:** No pasa nada. Las letras azules siguen allí, esperando a que resuelvas el acertijo. ¿Qué intentas hacer ahora? (*Nothing happens. The blue letters are still there, waiting for you to solve the puzzle. What do you want to do now?*)
 - **Player utterance:** Cuál es el acertijo? (*What is the riddle?*)
- **Turn #8.** Although the LLM mentions “a major river in the region”, it does not *read* the riddle itself for the player.
 - **GM utterance:** No pasa nada... El acertijo es un misterio que debes resolver para desbloquear el pasaje a la Celda. No te puedo dar la respuesta, pero puedo decirte que está relacionado con un río importante en la región (*Nothing happens... The riddle is a mystery that you must solve to unlock the passage to the Cell. I can't give you the answer, but I can tell you that it is related to a major river in the region.*)
 - **Player utterance:** Regreso camino al Claro en el monte en busca de más pistas (*I head back to the Clearing in the woods in search for more clues*)

This is the last page.

<https://www.pedeciba.edu.uy/es/area/informatica/>