



UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY



FACULTAD DE
INGENIERÍA

Evaluación Automática de la Oralidad en Aprendizaje de Inglés

Informe de Proyecto de Grado presentado por

Ana Clara Nóbile, Ignacio Viscardi

en cumplimiento parcial de los requerimientos para la graduación de las
carreras de Ingeniería Eléctrica e Ingeniería en Computación de la Facultad de
Ingeniería de la Universidad de la República

Supervisores

Pablo Cancela
Germán Capdehourat
Luis Chiruzzo
Santiago Góngora

Tribunal

Emilio Martínez
Lara Raad
Aiala Rosá

Montevideo, 25 de mayo de 2025



Evaluación Automática de la Oralidad en Aprendizaje de Inglés por Ana Clara Nóbile, Ignacio Viscardi tiene licencia [CC Atribución 4.0](#).

Agradecimientos

Queremos expresar nuestro más profundo agradecimiento a nuestras familias y amigos, quienes nos acompañaron incondicionalmente durante todos estos años de carrera. Su apoyo constante fue clave para que este proyecto y nuestra formación fueran posibles.

A nuestros tutores, Pablo, Germán, Luis y Santiago, por su invaluable guía y paciencia a lo largo de todo este proceso. Sus conocimientos y orientaciones fueron fundamentales para el desarrollo del proyecto.

Al programa *Ceibal en Inglés*, cuya motivación fue el origen de este proyecto. En particular, a Cecilia y Alina, por su disposición y genuino interés en la realización.

Por último, queremos reconocer a todas las personas que, de una u otra forma, colaboraron o formaron parte de este proceso, ya sea escuchándonos, compartiendo ideas o simplemente estando presentes.

Resumen

Este proyecto aborda el desafío de **automatizar la evaluación del habla en inglés como segunda lengua**, motivado por el interés del sistema educativo público en evaluar a una mayor cantidad de estudiantes de educación primaria y media en el país e impulsado por el programa educativo *Ceibal en Inglés* de Ceibal. El trabajo se centra en desarrollar y evaluar herramientas que permitan analizar tanto la forma de producción del habla como el contenido semántico de las respuestas orales.

Para la **evaluación de características acústicas**, se implementaron dos enfoques: una línea base que compara secuencias fonéticas utilizando el modelo *Wav2Vec2Phonemes* para la extracción de fonemas, y la adaptación del sistema *MultiPA* (*Multi-task Pronunciation Assessment*), un modelo preentrenado que evalúa múltiples aspectos de pronunciación. Estos métodos se validaron utilizando dos conjuntos de datos: *EpaDB* (*English Pronunciation by Argentinians Database*) y datos reales de estudiantes de Ceibal.

En el **análisis del contenido semántico**, se desarrolló un sistema basado en el modelo de lenguaje de gran escala *Llama 3* de 8 mil millones de parámetros. Se diseñaron y evaluaron diversos métodos de instrucción (*prompting*) para analizar el cumplimiento de tareas en cuatro tipos diferentes de ejercicios de oralidad. Para la transcripción del habla se utilizó *WhisperX*, seleccionado tras una comparativa con otros sistemas de reconocimiento automático del habla.

Los componentes desarrollados se integraron en una **plataforma web** unificada implementada con *React* en el *frontend* y *Python FastAPI* en el *backend*. La arquitectura modular del sistema permite la ejecución flexible de diferentes flujos de evaluación mediante un sistema configurable de *pipelines*, facilitando la incorporación de nuevas herramientas y métodos.

Los resultados muestran que *MultiPA* logra buenas correlaciones con evaluaciones humanas en términos de inteligibilidad y fluidez. En la evaluación semántica, las técnicas de *few-shot learning* y *chain of thought* resultaron particularmente efectivas, aunque se identificaron desafíos en la categorización de respuestas intermedias dentro de un esquema de tres niveles de corrección. La plataforma desarrollada muestra la viabilidad de integrar estas herramientas de manera funcional, estableciendo una base para futuras mejoras en la evaluación automatizada del habla en la enseñanza de idiomas.

Palabras clave: Evaluación automática del habla, Procesamiento de Lenguaje

Natural, Task Fulfillment, Evaluación de pronunciación, LLM

Índice general

1. Introducción	1
1.1. Motivación	1
1.2. Planteamiento del Problema	2
1.3. Objetivos	2
1.4. Justificación del Proyecto	2
1.5. Organización del Documento	3
2. Revisión de Antecedentes	5
2.1. Evaluación Automatizada en el Aprendizaje de un Segundo Idioma	5
2.2. Fundamentos de la Evaluación Automatizada del Habla	6
2.3. Prueba Actual de Oralidad de Ceibal en Inglés	8
2.4. Selección de Características a Evaluar	9
2.4.1. Consideraciones en el Caso de Estudio	9
3. Evaluación de Características Acústicas	11
3.1. Fundamentos de Pronunciación	11
3.1.1. Categorización de Técnicas de Evaluación Automática	13
3.1.2. Métodos de Evaluación de Pronunciación	15
3.2. Selección de Métodos	18
3.2.1. Bases de Datos	18
3.2.2. Disponibilidad de Código con Capacidad de Inferencia	22
3.2.3. Implementación de Métodos	22
3.3. Evaluación de Pronunciación	26
3.3.1. Evaluación de la Línea Base	26
3.3.2. Evaluación de MultiPA	31
3.4. Conclusiones del Capítulo	49
4. Evaluación de Características Semánticas	51
4.1. Introducción a la Evaluación Semántica	51
4.2. <i>Task Fulfillment</i>	52
4.2.1. Definición e Importancia	52
4.2.2. Categorización de Tareas	52
4.3. Base de Datos para la Evaluación de Task Fulfillment	53
4.4. Arquitectura del Sistema de Evaluación Propuesto	56

4.4.1.	Visión General del Sistema	56
4.5.	Reconocimiento Automático del Habla	58
4.5.1.	Comparativa de Sistemas ASR	59
4.5.2.	Resultados y Selección del ASR	60
4.6.	Analizador Semántico: Fundamentos y Diseño	62
4.6.1.	Arquitectura Interna del Analizador Semántico	62
4.6.2.	Modelos de Lenguaje de Gran Escala (LLMs) en Evaluación Semántica	63
4.6.3.	Diseño e Implementación del Analizador Semántico	66
4.6.4.	Reflexión general sobre el Diseño de <i>Prompts</i>	71
4.7.	Metodología de Evaluación	72
4.7.1.	Métricas	73
4.8.	Resultados y Análisis	74
4.8.1.	Análisis para Tareas de Tipo 1	74
4.8.2.	Análisis para Tareas de Tipo 2	86
4.8.3.	Análisis para Tareas de Tipo 3	92
4.8.4.	Análisis para Tareas de Tipo 4	96
4.8.5.	Análisis de Consistencia	103
4.9.	Conclusiones del Capítulo	105
4.9.1.	Líneas a Explorar	106
5.	Módulo Completo	109
5.1.	Introducción	109
5.2.	Arquitectura del Sistema	110
5.2.1.	Visión General	110
5.2.2.	Capa de Presentación (frontend)	111
5.2.3.	Capa de Lógica de Negocio (backend)	114
5.2.4.	Módulo Audiolang	117
5.3.	Flujo al evaluar una tarea	121
5.4.	Conclusiones del Capítulo	122
6.	Conclusiones y Trabajo Futuro	125
A.	Anexo 1	135
A.1.	Estadísticas Descriptivas de la Evaluación de <i>EpaDB</i> en <i>MultiPA</i>	135
A.2.	Estadísticas Descriptivas de la Evaluación de la Base de <i>Ceibal</i> en <i>MultiPA</i>	136
A.2.1.	Basadas en la Media	136
A.2.2.	Basadas en la Mediana	138
A.3.	PCA en Base de Ceibal con MultiPA	139
A.4.	Prompts para Evaluación de <i>Task Fulfillment</i>	139
A.4.1.	Prompts para Tareas de Tipo 1	139
A.4.2.	Prompts para Tareas de Tipo 2	145
A.4.3.	Tarea de Tipo 3	152
A.4.4.	Tarea de Tipo 4	154
A.5.	Análisis de Consistencia	156

Capítulo 1

Introducción

1.1. Motivación

En 2012 surge el programa *Ceibal en Inglés* a partir de la necesidad de universalizar la enseñanza de inglés a nivel de educación pública primaria en Uruguay. Mediante una solución de videoconferencia, se implementaron clases semanales impartidas por docentes remotos a estudiantes de diversos centros educativos del país.



Figura 1.1: Clase por videoconferencia de *Ceibal en Inglés*.

Actualmente, el programa se ha expandido para incluir también centros de enseñanza media pública, además de ser desde hace unos años quien cubre la mayor cantidad de estudiantes a nivel de primaria. El alcance nacional del programa plantea desafíos determinados por la cantidad de centros educativos y estudiantes a cubrir, especialmente en la **evaluación oral del aprendizaje**. Desde 2014, se ha implementado una prueba adaptativa informatizada para evaluar el nivel alcanzado por los estudiantes a medida que avanzan en el programa. Sin embargo, esta evaluación presenta una carencia importante al no incluir la evaluación oral del segundo idioma, lo cual es una de las habilidades más com-

plejas de evaluar debido a la alta carga de recursos humanos especializados que implica la realización de dichas pruebas.

Para abordar esta limitación, se desarrolló recientemente una evaluación de oralidad mediada por tecnología y de forma asincrónica, utilizando una plataforma en línea que evita la necesidad de interacciones cara a cara entre estudiantes y evaluadores. La prueba consiste en que los estudiantes respondan de forma oral a distintas consignas mediante grabaciones de audio, que luego son escuchadas y calificadas por evaluadores humanos expertos. Esta modalidad fue validada mediante estudios realizados en 2021 y 2022, en colaboración con el *Centre for Research in English Language Learning and Assessment (CRELLA)* de la *University of Bedfordshire*, que no encontraron diferencias significativas con las evaluaciones tradicionales. Si bien esta prueba mejora aspectos relacionados con la escalabilidad, evitando por ejemplo el traslado de evaluadores para realizar pruebas presenciales o videoconferencias uno a uno, sigue siendo una limitante contar con los recursos humanos necesarios para alcanzar cobertura nacional en todos los niveles.

1.2. Planteamiento del Problema

La necesidad de evaluar oralmente a un gran número de estudiantes en todo el país pone de manifiesto las limitaciones de los métodos tradicionales de evaluación, incluso aquellos mediados por tecnología. Esto plantea la pregunta:

¿Cómo podemos utilizar herramientas tecnológicas avanzadas para automatizar la evaluación del habla en inglés como segunda lengua, mejorando la escalabilidad y eficiencia del proceso?

1.3. Objetivos

El presente proyecto tiene como objetivo general explorar, implementar y evaluar herramientas de procesamiento de audio y lenguaje natural para automatizar la evaluación del habla en inglés como segunda lengua, con foco en el análisis de características acústicas y semánticas de las producciones orales.

De forma específica, el trabajo busca investigar las técnicas actuales de evaluación automatizada del habla y su aplicación en el ámbito educativo; diseñar e implementar métodos que permitan evaluar aspectos relacionados con la forma de producción del habla, así como con la coherencia y adecuación de las respuestas; y validar dichas herramientas mediante su aplicación en bases de datos relevantes, comparando los resultados obtenidos con las evaluaciones realizadas por evaluadores humanos.

1.4. Justificación del Proyecto

La implementación de herramientas automatizadas de evaluación del habla tiene el potencial de transformar la forma en que se evalúa la oralidad en el

aprendizaje de idiomas. Al abordar los desafíos actuales de escalabilidad y recursos, este proyecto busca sentar las bases para soluciones más eficientes y accesibles en la educación pública.

1.5. Organización del Documento

El presente documento se organiza en seis capítulos, que presentan un enfoque basado en herramientas automáticas para la evaluación de la oralidad en el aprendizaje del inglés como segunda lengua.

En el [Capítulo 2](#), se presenta una revisión detallada de los antecedentes necesarios para contextualizar el estudio. Se comienza explorando el campo del aprendizaje de idiomas asistido por computadora (*Computer-Assisted Language Learning, CALL*), su evolución histórica y su relevancia en la enseñanza de segundas lenguas. A continuación, se profundiza en los fundamentos de la evaluación automatizada del habla, abordando los desafíos en la estandarización de estos procesos. Finalmente, se describe la prueba actual de oralidad de *Ceibal en Inglés* y se justifica la selección de las características a evaluar en este estudio, estableciendo el marco sobre el cual se desarrollan los enfoques abordados en los capítulos siguientes.

El [Capítulo 3](#) se centra en la evaluación de las **características acústicas** del habla. Se revisan los fundamentos teóricos sobre pronunciación y se analiza cómo la percepción de lo que constituye una buena pronunciación ha evolucionado, pasando de un énfasis en sonar como un hablante nativo a priorizar la inteligibilidad y la comprensibilidad. Se clasifican y analizan diversas técnicas de evaluación automática, implementando dos enfoques principales: una línea base basada en la comparación fonética y el modelo avanzado *MultiPA*. Luego de una búsqueda exhaustiva de bases de datos adecuadas al público objetivo, se seleccionaron corpus pertinentes sobre los cuales se aplicaron y evaluaron los métodos. Los resultados evidencian las limitaciones de la línea base y destacan la eficacia de *MultiPA* en contextos donde el objetivo es mejorar la comunicación efectiva, especialmente en la categoría de *Delivery*, en el corpus de *Ceibal en Inglés*.

En el [Capítulo 4](#), se aborda la evaluación de las **características semánticas** de las respuestas orales. Se diseña y evalúa un componente capaz de analizar el cumplimiento de tareas utilizando herramientas de reconocimiento automático del habla y modelos de lenguaje a gran escala con técnicas de *prompting*. Se construye una base de datos representativa para validar diferentes *prompts* en cuatro tipos de tareas: preguntas simples con respuestas breves, preguntas basadas en imágenes, respuestas elaboradas que abordan criterios específicos e identificación de diferencias entre imágenes. Se analizan los desafíos inherentes a cada tipo de tarea y se evalúa el desempeño del modelo, identificando áreas de mejora y proponiendo refinamientos en los *prompts* y en el enfoque de evaluación.

El [Capítulo 5](#) integra todos los componentes desarrollados en una **plataforma web** unificada y funcional. Se detalla la arquitectura del sistema, basada en

una separación clara entre el *frontend*, el *backend* y el módulo de procesamiento de audio (*Audiolang*). Se utiliza *React* para el *frontend*, *FastAPI* en Python para el *backend* y *SQLite* como sistema de gestión de base de datos. Se describe la implementación de los componentes, el flujo de trabajo al evaluar una tarea y cómo la plataforma permite interactuar de manera sencilla con las herramientas desarrolladas. Se valida el funcionamiento de la plataforma tanto en entornos locales como en situaciones más desafiantes, lo que sugiere un buen grado de flexibilidad y adaptabilidad ante limitaciones de hardware y ambientes ruidosos.

Finalmente, el [Capítulo 6](#) presenta las conclusiones generales del trabajo, discute las limitaciones identificadas y propone direcciones para investigaciones futuras. Se resalta cómo la implementación de *MultiPA* y las técnicas de *prompting* con modelos de lenguaje a gran escala contribuyen a la evaluación automática de características acústicas y semánticas de respuestas orales. Se proponen líneas de trabajo futuro que incluyen la ampliación y etiquetado del conjunto de datos con contexto educativo real, el refinamiento de los *prompts*, el análisis de consistencia y la incorporación de modelos más avanzados que puedan procesar texto, imágenes y audio. Estas direcciones buscan mejorar la precisión, consistencia y utilidad del sistema desarrollado, adaptándolo mejor a las necesidades del entorno educativo y potenciando sus capacidades en programas como *Ceibal en Inglés*.

Capítulo 2

Revisión de Antecedentes

En este capítulo se presenta una revisión de los antecedentes necesarios para contextualizar el estudio de la evaluación automática de la oralidad en el aprendizaje de inglés. Se comienza explorando el campo del aprendizaje de idiomas asistido por computadora (*Computer-Assisted Language Learning, CALL*), su evolución histórica y su relevancia en la enseñanza de segundas lenguas. A continuación, se profundiza en los fundamentos de la evaluación automatizada del habla, abordando las características evaluadas según el *Marco Común Europeo de Referencia para las Lenguas* [1] y los desafíos en la estandarización de estos procesos. Finalmente, se describe la prueba actual de oralidad de *Ceibal en Inglés* y se justifica la selección de las características a evaluar en este estudio, lo que permite contextualizar el desarrollo de los métodos abordados posteriormente.

2.1. Evaluación Automatizada en el Aprendizaje de un Segundo Idioma

La evaluación automatizada del habla en el aprendizaje de idiomas se enmarca dentro del campo más amplio del aprendizaje de idiomas asistido por computadora, definido como “la búsqueda y el estudio de aplicaciones del ordenador en la enseñanza y el aprendizaje de idiomas” [2].

Inicialmente, en las décadas del 60 y 70, estos sistemas se centraban en la enseñanza basada en ejercicios repetitivos y estructurados, como la práctica de gramática y vocabulario a través de respuestas predeterminadas. Luego se introdujeron actividades más flexibles, permitiendo participar en simulaciones, juegos y otros ejercicios interactivos con retroalimentación instantánea.

Para dar una mayor referencia de lo que es un *CALL* podemos pensar en las aplicaciones comerciales conocidas como *Duolingo*, *ELSA Speak* o *Voxy*, que tienen entornos interactivos para el aprendizaje de idiomas. En particular, mencionamos estas tres ya que evalúan aspectos de la oralidad. *Duolingo* en su entorno tiene una sección especialmente dedicada a mejorar la pronunciación,

en la cual el usuario debe repetir palabras o frases definidas y son evaluadas de forma automática [3]. *Voxy* tiene un examen para determinar el nivel de inglés actual, llamado *Voxy Proficiency Assessment (VPA)*, el cual tiene una sección de oralidad. En este caso, la evaluación es efectuada de forma manual por especialistas, tal como se describe en su página oficial [4].

Por su parte, *Elsa Speak* se centra especialmente en la evaluación de la oralidad. Tiene una herramienta, llamada *Speech Analyzer*, la cual a partir de ejercicios del tipo lectura de un texto y respuesta espontánea a una consigna, evalúa de forma automática la pronunciación, entonación, fluidez, gramática y vocabulario [5].

De todas formas, hay muy poca información disponible sobre el diseño de estas aplicaciones comerciales y los métodos que utilizan para evaluar, tal como se menciona en la sección 2.1.2 del libro *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech* [6].

La evaluación automática de la oralidad es un componente actualmente presente en los sistemas CALL y se enmarca dentro de los ASA (Automatic Speech Assessment). Esta evaluación se basa tanto en la forma de producción del habla como en el contenido y la coherencia de sus respuestas. En las siguientes secciones, profundizaremos en esta distinción.

2.2. Fundamentos de la Evaluación Automatizada del Habla

El *Marco Común Europeo de Referencia para las Lenguas (CEFR - Common European Framework of Reference for Languages)* presenta una escala global de evaluación de un idioma compuesta por 6 niveles: A1, A2, B1, B2, C1, C2. Respecto a la evaluación de la oralidad, la divide en 5 características distintas: *Range*, *Accuracy*, *Fluency*, *Interaction* y *Coherence*, las cuales a su vez también son clasificadas en los 6 niveles mencionados anteriormente [1].

Cada una de estas características puede ser evaluada de forma independiente, lo que se conoce como *single-view grading*. El enfoque *single-view grading* es lo que generalmente se automatiza porque permite a los sistemas enfocarse en un solo aspecto del rendimiento del hablante, analizando parámetros específicos. De todas formas, transformar la descripción dada por el CEFR a características medibles no es un proceso directo.

Existen vastos estudios sobre la evaluación automática de la oralidad, pero no todos siguen el mismo criterio respecto a qué características es necesario evaluar ni cuáles son las métricas asociadas a cada una. Tampoco todos consideran las escalas mencionadas por el CEFR. Es decir, la evaluación automática de la oralidad no es un proceso estandarizado y varía según el enfoque utilizado, el objetivo de la evaluación y las herramientas empleadas. Esta diversidad en los métodos y criterios de evaluación hace que no exista una única forma de medir los aspectos relacionados a la oralidad, lo que puede llevar a resultados dispares entre distintos sistemas.

Es por esto que la automatización de la evaluación plantea desafíos no solo técnicos, sino también conceptuales. Esto también implica que los resultados obtenidos por distintos sistemas no siempre son equivalentes, lo que subraya la necesidad de una mayor estandarización en este campo.

Por nuestra parte y al igual que en varios enfoques, en particular el de Kyriakopoulos en *Deep Learning for Automatic Assessment and Feedback of Spoken English* [7], separaremos la evaluación en dos características fundamentales: las asociadas a la **forma de producción** del habla y las asociadas al **contenido y coherencia** de una respuesta.

Las características asociadas a la **forma de producción** del habla generalmente evalúan aspectos como la pronunciación, fluidez y prosodia. De todas formas, estas tres características no siempre se categorizan como independientes. Diversos enfoques sostienen que la fluidez y prosodia son aspectos inherentes a la pronunciación, y a su vez existen distintos criterios respecto a lo que es una correcta pronunciación y cómo medirla (una mayor profundización se encuentra en la sección 3.1). A su vez, hay otros conceptos asociados a estas características tales como el tiempo, estrés, ritmo y entonación.

Estos enfoques utilizan distintas técnicas para evaluar cada una de las características mencionadas. Para la evaluación de prosodia generalmente se consideran tres aspectos principales: la frecuencia fundamental (F0), que mide los cambios en la altura del sonido e identifica la entonación; la energía, que evalúa la intensidad o volumen de las sílabas; y la duración, que analiza el tiempo empleado en la producción de fonemas y las pausas [8]. De todas formas, cómo estas tres dimensiones influyen en la prosodia puede ser subjetivo.

Entre las características generalmente consideradas para la evaluación de fluidez se encuentran el análisis de palabras por minuto (*Words Per Minute - WPM*), la duración promedio del habla entre pausas y el porcentaje de pausas en el tiempo total del habla [9]. Para el caso particular de habla leída, se tiene en cuenta el número de palabras correctas por minuto (*Words Correct Per Minute - WCPM*) [10].

La pronunciación es el método más estudiado académicamente y el más amplio, ya que algunos estudios consideran que engloba a las otras características. Hay diversos enfoques para la evaluación de pronunciación, que van desde una comparación fonética hasta métodos de confianza derivados de sistemas de reconocimiento automático del habla. Una profundización sobre los distintos métodos de evaluación de pronunciación se abordan en la sección 3.1.2.

Las características asociadas al **contenido y coherencia** incluyen aspectos como la riqueza léxica, la corrección gramatical, la coherencia semántica y la organización lógica del discurso. La riqueza léxica se refiere al uso variado y preciso del vocabulario; la corrección gramatical se enfoca en la adecuación de las estructuras sintácticas. La coherencia semántica evalúa si la respuesta es pertinente en relación con la pregunta formulada, y la organización lógica del discurso analiza cómo se estructuran y conectan las ideas para mantener un flujo coherente.

El enfoque generalmente utilizado para la evaluación automática de estas características se basa en utilizar un sistema de reconocimiento automático del habla (ASR, por sus siglas en inglés) para transcribir el discurso y luego analizar el texto obtenido. Esto tiene el desafío inherente de depender completamente de la precisión de la transcripción del ASR. Si el ASR comete errores (por ejemplo, al reconocer incorrectamente palabras, omitir términos clave o malinterpretar disfluencias comunes del habla), estos fallos se trasladan a la evaluación, lo que puede llevar a un resultado inexacto. Esto hace que la calidad de la evaluación final esté directamente ligada a la fiabilidad del ASR.

2.3. Prueba Actual de Oralidad de Ceibal en Inglés

En la actualidad, la prueba de oralidad que realiza *Ceibal en Inglés* se lleva a cabo mediante una plataforma web y está compuesta por cuatro partes, con distinta cantidad de consignas del mismo tipo dentro de cada una [11].

A continuación, se describen brevemente las partes y los tipos de consignas que componen la evaluación:

- **Preguntas simples con respuesta libre:** Consisten en consultas básicas en inglés, como “What’s your name?” o “Can you swim?”
- **Preguntas contextuales basadas en imágenes:** Involucran la observación de una imagen y la respuesta a una pregunta relacionada, como una imagen de unos lentes acompañada de la pregunta “What are they?”
- **Tareas guiadas con respuesta elaborada:** Requieren la producción de un discurso idealmente más extenso y estructurado, por ejemplo, “Tell me about your favourite famous person”.
- **Tareas de comparación y contraste:** Implican la comparación de dos láminas y la descripción de sus diferencias.

En cada parte se plantean distintas consignas y se le da determinada cantidad de tiempo al estudiante para que grabe su respuesta. En total, la prueba consiste en 14 consignas distintas distribuidas entre las 4 partes de la prueba.

Luego, evaluadores expertos califican las respuestas de forma holística en tres categorías: Task Fulfilment, Language (Range & Accuracy) y Delivery (Intelligibility & Flow of Speech). El estudiante recibe un único puntaje global en cada categoría (de 0 a 3), en lugar de calificaciones individuales por respuesta. Se considera como puntaje total de la prueba la suma de los puntajes en estas tres categorías.

Los puntajes del 1 al 3 corresponden a los niveles *Pre-A1*, *A1* y *A2* respectivamente. El puntaje 0 se otorga cuando la respuesta no puede ser evaluada,

lo cual le otorga la categoría *Not rateable*. Este caso ocurre cuando no hay contenido en inglés en los audios, está limitado a la lectura de las consignas o el evaluador identificó que el estudiante tuvo ayuda externa al realizar su respuesta.

2.4. Selección de Características a Evaluar

Las categorías de evaluación de la prueba de *Ceibal en Inglés* se pueden clasificar según los dos criterios mencionados en la sección 2.2: las asociadas a la **forma de producción del habla**, y las asociadas al **contenido y coherencia** de una respuesta.

Dentro de las de contenido y coherencia se encuentran *Task Fulfillment* y *Language*, mientras que *Delivery* está asociada a la forma de producción.

Se centrará la automatización en la evaluación de dos de estas características: *Task Fulfillment* y *Delivery*.

La característica *Task Fulfillment* se basa en evaluar la correctitud de la respuesta respecto a la pregunta o consigna planteada. El nivel A2 respecto a esta categoría ocurre cuando el estudiante responde correctamente a las 14 consignas de la prueba y el significado de su respuesta se comunica de forma satisfactoria. El nivel A1 cuando responde correctamente al menos a 7, y Pre-A1 si son al menos 3. En el capítulo 4.1 se abordará la evaluación de esta característica adoptando un enfoque por tarea individual, en contraposición a una evaluación global de la prueba completa. A la evaluación de las características que hacen el *Task Fulfillment*, asociadas con el contenido y coherencia de una respuesta, se les denominará **características semánticas**.

Para la evaluación de la categoría *Delivery*, un desempeño correspondiente al nivel A2 implica que “casi todas las palabras o frases son inteligibles” y que el hablante “logra mantener el discurso por un período, aunque persisten las pausas”. Esta categoría integra tanto *Intelligibility* como *Flow of Speech*, por lo que se decide abordar su evaluación automática mediante el estudio de características asociadas a la pronunciación. Esto es debido a que académicamente estas características han sido ampliamente estudiadas para evaluar de forma automática y porque la pronunciación en sí también contempla la inteligibilidad en una de sus definiciones. Una profundización sobre esto se encuentra en la sección 3.1. A estas características, asociadas con la forma de producción del habla, se les denominará **características acústicas**.

2.4.1. Consideraciones en el Caso de Estudio

Generalmente, la evaluación de características acústicas y semánticas en el habla incluye un módulo ASR en alguna de sus etapas. Los modelos más recientes han mostrado un buen desempeño en general, pero es necesario evaluar su

eficacia con hablantes de inglés no adultos y con acento español rioplatense.

Usualmente, los sistemas ASR están entrenados con grandes conjuntos de datos de habla de adultos nativos (en su mayoría provenientes de contextos formales, como entrevistas o lecturas en voz alta). Por lo tanto, suelen tener dificultades para entender el habla de personas no nativas, debido a la falta de representatividad en los conjuntos de entrenamiento y construcciones no gramaticales, influenciados en gran medida por la lengua materna y el general bajo nivel del hablante [12]. Aunque los últimos avances han ampliado la diversidad en los datos de entrenamiento, es importante analizar su rendimiento en estos contextos específicos, ya que la mayoría de las mejoras suelen enfocarse en idiomas o acentos más globalmente hablados o conocidos, lo que no necesariamente incluye variaciones como el español rioplatense.

Por otra parte, el habla en niños y jóvenes presenta desafíos para los sistemas ASR también debido a la falta de representatividad en el entrenamiento de los modelos, a su variabilidad, a las diferencias en el desarrollo lingüístico y a la frecuente aparición de disfluencias. Además, las diferencias fisiológicas, como la menor longitud del tracto vocal, complican la alineación con modelos entrenados principalmente con datos de adultos [13].

Dado que estos modelos se utilizan tanto para la evaluación semántica como acústica, las consideraciones mencionadas anteriormente son inherentes a todo el estudio. La validación de las herramientas construidas estará condicionada por la disponibilidad de bases de datos que permitan verificar las características mencionadas.

En los próximos capítulos, se abordará la evaluación de las dos dimensiones del habla: las características acústicas y las características semánticas. Cada una de estas secciones incluirá la fundamentación teórica necesaria, una revisión de antecedentes, el diseño de la metodología empleada, así como la evaluación correspondiente utilizando las bases de datos seleccionadas.

Capítulo 3

Evaluación de Características Acústicas

En este capítulo se revisan los fundamentos teóricos sobre pronunciación, junto con una clasificación y análisis de las técnicas de evaluación automática. Se seleccionaron métodos específicos para el análisis, y se llevó a cabo una búsqueda exhaustiva de bases de datos adecuadas al público objetivo, en las cuales se aplicaron y evaluaron algunos de estos métodos.

3.1. Fundamentos de Pronunciación

La pronunciación es un aspecto esencial en la adquisición de una segunda lengua, ya que influye directamente en la capacidad de un hablante para comunicarse de manera efectiva y ser comprendido. A lo largo del tiempo, las definiciones de lo que constituye una “buena” pronunciación han variado. Tradicionalmente, se consideraba que la buena pronunciación estaba estrechamente vinculada con la habilidad de un hablante para sonar como un hablante nativo, un enfoque predominante en la década de 1960.

Una de las propuestas más influyentes dentro del enfoque tradicional fue la de Robert Lado, quien proponía que la enseñanza de la pronunciación debía centrarse en la reproducción fiel de los sonidos del idioma objetivo, tanto a nivel segmental (fonemas) como suprasegmental (entonación, ritmo, acento) [14]. La meta era alcanzar una pronunciación lo más cercana posible a la de un hablante nativo, con especial énfasis en minimizar el acento extranjero. Este enfoque estaba alineado con la *teoría del contraste* [15], que identificaba las diferencias entre los sistemas fonológicos de la lengua materna y la segunda lengua como una fuente de dificultades en la pronunciación. En este contexto, Lado introdujo el *método de puntos discretos*, que descomponía el idioma en pequeños segmentos evaluables, permitiendo identificar y abordar los problemas específicos derivados de esas diferencias fonológicas [16].

Hacia finales del siglo XX, el enfoque tradicional de la enseñanza de la pronunciación comenzó a cuestionarse, impulsado por el trabajo de Munro y Derwing [17]. Estos autores propusieron una perspectiva más orientada a la comunicación, que no se centraba exclusivamente en sonar como un hablante nativo, sino en la comprensibilidad e inteligibilidad del discurso. En su estudio *Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners*, los autores comenzaron a desenredar conceptos relacionados, pero parcialmente independientes, como el acento extranjero (*accentedness*), la inteligibilidad (*intelligibility*) y la comprensibilidad (*comprehensibility*).

Inteligibilidad se refiere al grado en que el mensaje de un hablante es comprendido correctamente por el oyente. **Comprensibilidad**, por su parte, está relacionada con el esfuerzo que un oyente debe realizar para entender el discurso, ya que un enunciado puede ser inteligible pero requerir un mayor esfuerzo para ser procesado. Por último, el **acento** describe la medida en que la pronunciación de un hablante difiere de las normas de un hablante nativo, sin que esto implique necesariamente una barrera para la inteligibilidad o la comprensibilidad.

Hablar una segunda lengua con acento extranjero no necesariamente impide la comunicación efectiva, aunque ciertas características pueden complicarla. Munro y Derwing sostienen que el principal objetivo de la enseñanza de la pronunciación debe ser asegurar que los hablantes sean comprensibles, es decir, que su mensaje pueda ser entendido claramente, sin que sea necesario sonar como un hablante nativo. Este enfoque sigue el principio de inteligibilidad formalizado por John M. Levis [18], que prioriza la comprensión del mensaje por sobre la imitación de una pronunciación nativa, en contraposición al principio de natividad, que promueve la idea de alcanzar una pronunciación lo más cercana posible a la de un hablante nativo [19].

Aunque hay una creciente cantidad de estudios que respaldan el enfoque de la inteligibilidad como prioridad en la enseñanza de la pronunciación [20], tanto las prácticas docentes como la mayoría de los estudios empíricos en pronunciación siguen basándose en el principio de natividad [21]. Esta contradicción puede estar relacionada con el hecho de que el campo de estudio es relativamente joven y que la diferencia entre los términos inteligibilidad, comprensibilidad y acento a veces no queda clara y tiende a confundir. No es así en el caso del programa *Ceibal en Inglés*, quienes actualmente sí evalúan inteligibilidad.

Es importante destacar que la inteligibilidad y la comprensibilidad son términos más asociados a cómo un humano interpreta el habla, lo que dificulta la tarea de modelar estas dimensiones de manera automática. Mientras que el acento puede identificarse de manera más directa a través de características acústicas específicas, la evaluación de la inteligibilidad y la comprensibilidad son dimensiones vinculadas a la interpretación subjetiva del oyente, lo que hace que sea complejo traducir estas definiciones humanas a métricas automáticas.

Esto lleva a la creación de una gran cantidad de enfoques distintos para la evaluación de dicha competencia, que se abordarán en la siguiente sección.

3.1.1. Categorización de Técnicas de Evaluación Automática

Tras analizar las diversas perspectivas sobre lo que constituye una buena pronunciación (desde sonar como un hablante nativo hasta priorizar la inteligibilidad y comprensibilidad) es necesario explorar cómo se evalúa esta habilidad en el contexto de las herramientas de procesamiento del habla. La evolución en la definición de pronunciación ha llevado a una reevaluación de los métodos utilizados para su análisis. En lugar de buscar una pronunciación “correcta” o “incorrecta” en términos absolutos, se reconoce que la pronunciación se sitúa en un espectro que va desde lo ininteligible hasta el habla similar a la de un nativo [22]. Esta comprensión más matizada lleva al desarrollo de enfoques de evaluación que consideran múltiples dimensiones del habla, reflejando la complejidad inherente a la comunicación oral.

Dado que cuantificar los errores de pronunciación es una tarea compleja, estos se clasifican a menudo en **evaluaciones objetivas**, que abarcan aspectos fonéticos, prosódicos y articulatorios, y **evaluaciones subjetivas**, centradas en la inteligibilidad, comprensibilidad y acento [23]. Esta clasificación permite abordar la pronunciación desde diferentes ángulos, reconociendo tanto las características acústicas medibles como las percepciones del oyente. En este contexto, han surgido diversos enfoques para evaluar la competencia en pronunciación, cada uno enfocado en distintas características, como se ilustra en la siguiente figura:

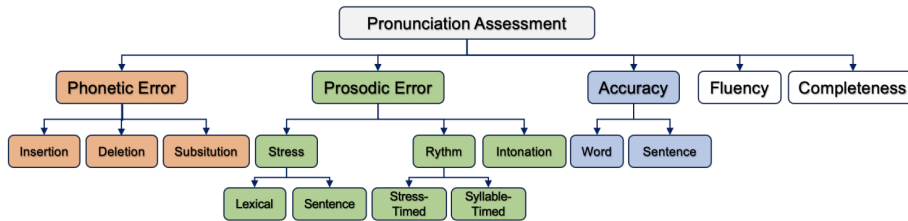


Figura 3.1: Tipos de errores de pronunciación para evaluación. Extraída de [23].

En la figura, la evaluación de la pronunciación se clasifica en cinco grupos, siendo los dos más grandes: errores fonéticos y errores prosódicos. Los errores fonéticos incluyen inserciones, eliminaciones y sustituciones de fonemas, centrándose en la precisión en la producción de los sonidos individuales. Los errores prosódicos abarcan elementos suprasegmentales como el estrés, el ritmo y la entonación, que influyen en la naturalidad del habla. En este contexto, estrés se refiere a la mayor prominencia que se da a ciertas sílabas dentro de una palabra o a ciertas palabras dentro de una oración. Ritmo alude al patrón temporal del habla, relacionado con la duración y alternancia de sílabas acentuadas y no acentuadas. Entonación, por su parte, se refiere al movimiento del tono (frecuencia fundamental) a lo largo del enunciado, que contribuye a expresar

significado, intención comunicativa y estructura discursiva.

Es importante destacar que la **fluidez** es una característica que algunos autores consideran separada de la pronunciación, enfocándose en aspectos como la velocidad, la duración de las pausas y la continuidad del discurso. Estos elementos afectan directamente la facilidad con la que el oyente puede comprender el habla, independientemente de la precisión fonética o prosódica. Sin embargo, otros investigadores integran la fluidez dentro del ámbito de la pronunciación, argumentando que está intrínsecamente ligada a la producción efectiva del habla y a la implementación adecuada de los patrones prosódicos. Esta dualidad en la consideración de la fluidez refleja las diversas perspectivas en el campo del procesamiento del habla y subraya la complejidad de evaluar competencias orales.

Además, se consideran otras dimensiones importantes como la precisión (*accuracy*) y la completitud (*completeness*), que contribuyen a una evaluación integral de la pronunciación. La precisión se refiere a la exactitud en la articulación de los sonidos y patrones esperados, mientras que la completitud evalúa si todos los elementos lingüísticos requeridos están presentes en el discurso del hablante.

Kyriakopoulos, en su tesis [7], presenta una clasificación de los métodos de evaluación de pronunciación en las siguientes categorías:

- **Métodos de comparación con hablantes nativos (Native Speaker Comparison Methods):** Evalúan la similitud entre la pronunciación del estudiante y la de un hablante nativo, midiendo la proximidad entre características acústicas directamente obtenidas de los audios.
- **Métodos de medida de confianza (Confidence Measure Methods):** Evalúan la inteligibilidad del discurso, es decir, cuán comprensible es para un oyente o un sistema automático. Los sistemas de reconocimiento automático del habla (ASR) son una base común para estos métodos, ya que las palabras o fonemas reconocidos con baja confianza suelen corresponder a pronunciaciones incorrectas.
- **Métodos de reconocimiento canónico (Recognition Methods):** Determinan si el hablante se adhiere a una pronunciación estándar o canónica, detectando desviaciones de las normas establecidas para cada palabra o fonema.
- **Métodos supervisados (Supervised Methods):** No asumen reglas o estructuras predefinidas; en cambio, aprenden criterios de evaluación a partir de datos anotados por expertos humanos, lo que permite mayor flexibilidad y adaptación a las particularidades de los hablantes no nativos.

Además de las clasificaciones presentadas, los métodos de evaluación de pronunciación también se pueden categorizar según el origen de los datos utilizados para su entrenamiento y la modalidad del habla evaluada:

Según el **origen de los datos** utilizados para su entrenamiento se clasifican en:

- **Datos de hablantes nativos:** Datos de entrenamiento compuestos exclusivamente por grabaciones y transcripciones de hablantes nativos del idioma objetivo.
- **Datos de hablantes no nativos:** Datos de entrenamiento que incluyen grabaciones y transcripciones de hablantes no nativos, es decir, personas que están aprendiendo el idioma como segunda lengua.

Los métodos que utilizan exclusivamente datos de hablantes nativos para entrenar los modelos de evaluación pueden limitar la capacidad del sistema para reconocer y evaluar adecuadamente las variaciones comunes en la pronunciación de hablantes no nativos. Por otro lado, los métodos que incorporan datos de hablantes no nativos suelen ofrecer mejores resultados al evaluar a estudiantes de segundas lenguas, ya que el modelo está expuesto a la variabilidad típica de las pronunciaciones no nativas. Sin embargo, estos métodos enfrentan el desafío significativo de la escasez de datos etiquetados disponibles y la gran variabilidad en las pronunciaciones no nativas, lo que puede complicar el entrenamiento y la generalización del modelo [24].

La clasificación según la modalidad del habla evaluada presenta las siguientes categorías:

- **Habla leída:** Discurso producido al leer en voz alta un texto proporcionado previamente. El estudiante pronuncia textos predefinidos que han sido preparados de antemano. En la literatura esta clasificación también se conoce como “Respuesta cerrada” (*close response*).
- **Habla espontánea:** Enunciados que el hablante produce en respuesta a un estímulo o pregunta, sin leer ni recitar texto previamente preparado. En la literatura esta clasificación también se conoce como “Respuesta abierta” (*open response*).

La evaluación basada en el habla espontánea se considera más útil en la medida en que el discurso de entrada es más representativo del que se espera que produzca el hablante en la comunicación normal. Sin embargo, la evaluación del habla espontánea presenta desafíos adicionales debido a su naturaleza impredecible y variabilidad [7].

Por otro lado, el habla leída permite control sobre el contenido lingüístico y facilita la comparación directa con pronunciaciones estándar. Además permite la comparación directa entre estudiantes ya que se estaría evaluando a todos sobre las mismas frases.

3.1.2. Métodos de Evaluación de Pronunciación

Uno de los enfoques más aceptados y utilizados en la evaluación automática de pronunciación es el método *Goodness of Pronunciation* (GoP). Este método fue introducido por S.M Witt y S.J Young [25] y se basa en la estimación de

puntajes para cada fonema en una frase, comparando los fonemas producidos por el hablante con los fonemas esperados según la transcripción.

El GoP utiliza el concepto de *probabilidad posterior*, que es la probabilidad de que un evento haya ocurrido dado un conjunto de observaciones. En el caso del GoP, esta probabilidad posterior mide cuán probable es que el fonema producido por el hablante coincida con el fonema que debería haber sido pronunciado, dado el conjunto de características acústicas extraídas de la señal de audio.

El algoritmo GoP alinea el audio con su transcripción para identificar la ubicación exacta de cada fonema. A continuación, calcula un puntaje para cada uno de ellos basado en la probabilidad posterior de que el fonema producido corresponda al fonema objetivo. Finalmente, este puntaje se compara con un umbral predefinido para determinar si el fonema fue pronunciado correctamente o no.

Inicialmente, las probabilidades posteriores se calculaban mediante modelos de mezcla gaussiana y modelos ocultos de Markov (GMM-HMM). En investigaciones más recientes, se ha mejorado el método reemplazando los GMM-HMM por redes neuronales profundas (DNN-HMM), lo que ha dado mejores resultados en el cálculo de las probabilidades [26].

Sin embargo, a pesar de su amplia aceptación a nivel de academia, el método GoP presenta limitaciones que dificultan su uso en entornos prácticos. En general, no ofrece resultados suficientemente precisos para aplicaciones en campo, ya que tiene dificultades para identificar errores específicos de pronunciación, como omisiones, inserciones o sustituciones de fonemas.

En el estudio de Vidal et al. [27], titulado *Phone-Level Pronunciation Scoring for Spanish Speakers Learning English Using a GOP-DNN System*, se realizó inferencia en *GoP* con la base de datos *EpaDB (English Pronunciation by Argentinians Database)*, diseñada específicamente para la evaluación de la pronunciación de hablantes hispanohablantes de la región del Río de la Plata que están aprendiendo inglés como segunda lengua. Este estudio reportó que el desempeño del *GoP* fue inaceptablemente bajo para muchos fonemas, por lo que no se considera adecuado para aplicaciones en las que es crucial evitar frustrar a los estudiantes con intervenciones incorrectas frecuentes, limitando así su utilidad en entornos educativos prácticos.

Con el objetivo de superar las limitaciones del método, se han desarrollado variantes más recientes, como el *GoPT*, que incorpora una arquitectura basada en *Transformers*, mejorando la evaluación en la base de datos *speechocean762* [28] permitiendo un análisis más detallado de otros aspectos de la pronunciación a distintos niveles de granularidad [29].

Además, tal como se mencionó en la sección 2.4.1 la evaluación de la pronunciación en niños y adolescentes añade una complejidad adicional. Según Cao et al. [30], las variantes del GoP son menos confiables cuando se aplican a niños, debido al desarrollo incompleto del habla en edades tempranas y la falta de bases de datos adecuadas para este grupo etario.

Existen varios enfoques recientes en la evaluación automática de pronunciación. Los *modelos de extremo a extremo (End-to-End Models)*, que incluyen

arquitecturas como *LSTM-CTC*, han sido utilizados para la detección de errores de pronunciación sin la necesidad de alineaciones explícitas. Por ejemplo, Leung et al. [31] utilizaron una arquitectura *CNN-RNN-CTC* para evaluar la pronunciación a nivel de fonema.

Asimismo, los modelos auto-supervisados (*Self-Supervised Models*), como *wav2vec 2.0*, han sido aplicados con éxito en estudios recientes. Xu et al. [32] hicieron fine-tuning sobre *wav2vec 2.0* para la predicción de fonemas a nivel de frame, mientras que Kim et al. [33] no solo hicieron un fine-tuning sobre *wav2vec 2.0*, sino también sobre *HuBERT*, para predecir puntajes de prosodia y fluidez. Del mismo modo, *MultiPA*, basado en *HuBERT*, también emplea fine-tuning para la evaluación de múltiples aspectos de la pronunciación en escenarios de respuesta abierta y cerrada [34].

La Figura 3.2, extraída del paper *Automatic Pronunciation Assessment - A Review* de Yassine El Kleir, muestra un resumen del rendimiento de diferentes modelos de detección de pronunciación fonética evaluados en la base de datos *L2-ARCTIC*. En este gráfico, el rendimiento se mide utilizando la métrica *F1*, que calcula la media armónica entre precisión y exhaustividad.

Como se puede observar, los primeros modelos basados en *HMM-GMM* alcanzaban un rendimiento menor al 42 % en términos de la métrica *F1*, mientras que los enfoques más recientes, como los modelos multi-tarea y multi-modal, han elevado el rendimiento al rango del 61-63 %. Este progreso ha sido impulsado por el uso de técnicas avanzadas como los modelos de aprendizaje auto-supervisado y enfoques multi-linguales, mejorando la capacidad de los sistemas para capturar errores de pronunciación de manera más precisa en hablantes no nativos [34].

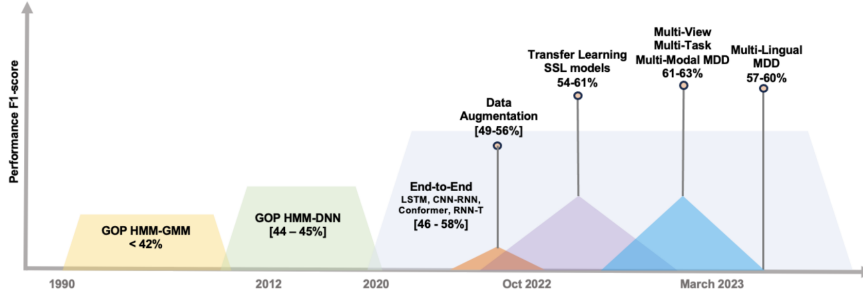


Figura 3.2: Resumen del rendimiento de diferentes modelos de detección de pronunciación fonética en L2-ARCTIC. Extraída de [23].

La Figura 3.3 muestra una evaluación del rendimiento de diferentes modelos en la detección de fluidez y prosodia, utilizando la base de datos *speechocean762*. Estas características, aunque a veces se evalúan por separado, también se consideran parte de la evaluación general de la pronunciación en muchos estudios como se mencionó en la sección 3.1.1.

El gráfico utiliza el coeficiente de correlación de Pearson como métrica pa-

ra medir el rendimiento en la evaluación de la fluidez y la prosodia. Como se observa, el rendimiento de los modelos ha mejorado de forma continua, comenzando con un rendimiento del 65.20 % y alcanzando entre 78-79 % con el uso de *MultiPA*.

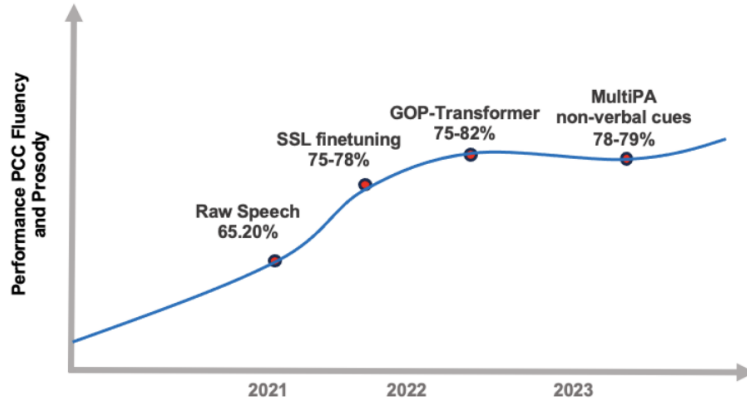


Figura 3.3: Resumen del rendimiento de diferentes modelos en la evaluación de fluidez y prosodia en la base *speechocean762*. Extraída de [23].

3.2. Selección de Métodos

La selección de los métodos de evaluación de pronunciación que se implementan en este trabajo se basa en dos criterios clave: las **bases de datos accesibles** y la **disponibilidad de código con capacidad de inferencia**. Estos factores se consideran fundamentales para evaluar la aplicabilidad práctica de los métodos seleccionados. A continuación, se detallan ambos criterios, que guían la elección de los métodos.

3.2.1. Bases de Datos

Para desarrollar o validar sistemas de evaluación de la pronunciación en hablantes no nativos, es necesario contar con datos anotados detalladamente a niveles de frase, palabra, fonema o la resolución a la que se pretenda evaluar.

Sin embargo, la mayoría de las investigaciones respecto a evaluación de pronunciación utilizan bases de datos privadas inaccesibles, lo que se traduce en una escasez notable de conjuntos de datos disponibles para investigación. Dentro de estos pocos conjuntos de datos accesibles, algunos tienen transcripción a nivel de fonema o palabra. Un número menor cuenta con evaluaciones manuales a nivel global, de frase, palabra o sobre la corrección en la pronunciación de fonemas específicos [35].

Es preferible que las bases de datos utilizadas incluyan hablantes nativos del idioma del grupo a evaluar, en este caso, el español. Esta especificidad garantiza una mayor correspondencia con los errores y características fonéticas propias de los hablantes de dicho idioma [36].

Dado que para este estudio el objetivo es evaluar la pronunciación en niños y adolescentes hispanohablantes, resulta indispensable contar con una base de datos que cumpla con estos requisitos, tanto para entrenar modelos como para evaluar su rendimiento. En la presente sección se resumen brevemente las bases de datos disponibles que presentan estas características.

Bases de datos públicas

Hasta la fecha, existen únicamente dos bases de datos de acceso público para evaluación de pronunciación con hablantes cuya lengua materna es español: “L2-ARCTIC: a non-native English speech corpus” y “EpaDB: a database of non native English speech by Spanish speakers from Argentina”, ambas centradas en evaluación a nivel de fonema. Por otro lado, también existe *speechocean762*, que, aunque está compuesta por hablantes cuya lengua materna es el mandarín, resulta útil presentarla debido a que la mitad de los hablantes son niños y, además, es un recurso ampliamente utilizado para entrenar algunos modelos.

L2-ARTIC

Se compone de 26,867 oraciones habladas por 24 hablantes no nativos, cuyos idiomas maternos (L1) incluyen hindi, coreano, mandarín, árabe, vietnamita y español. Cada idioma está representado por cuatro hablantes, dos hombres y dos mujeres.

Para cada hablante, se cuenta con aproximadamente una hora de lectura basada en *prompts de CMU ARCTIC*, un conjunto de oraciones diseñadas específicamente para la recopilación de datos de voz. Son fonéticamente equilibradas, lo que significa que están construidas para cubrir una gran cantidad de fonemas y combinaciones fonéticas comunes del idioma inglés.

Cada oración ha sido transcrita de forma automática tanto a nivel de palabra como fonema y alineados de forma forzada. Las etiquetas manuales que aporta esta base de datos son a nivel de fonema y tienen como fin identificar tres tipos de errores de pronunciación: sustituciones, eliminaciones y adiciones. Para la notación a nivel de fonema se utilizan símbolos *ARPAbet*. Además algunos archivos contienen comentarios brindados por los evaluadores, que pueden contener símbolos IPA.

ARPAbet y el Alfabeto Fonético Internacional (*IPA*) son sistemas complementarios para la representación de los sonidos del habla, diseñados con propósitos diferentes. *ARPAbet*, desarrollado en la década de 1970 por la Agencia de Proyectos de Investigación Avanzada (ARPA) [37], utiliza una codificación basada en caracteres ASCII para representar fonemas del inglés americano, lo que facilita su uso en aplicaciones computacionales. Por otro lado, el *IPA*, creado por la Asociación Fonética Internacional en el siglo XIX, es un sistema diseñado

para proporcionar un símbolo único para cada unidad distintiva del habla en cualquier idioma, permitiendo una transcripción uniforme y precisa [38].

El formato en el que se presentan las anotaciones es *TextGrid*, un estándar en la lingüística computacional. Dicho formato se utiliza en conjunto con herramientas como *Praat*, que permite la sincronización de estas anotaciones con la onda de voz y el espectrograma.

En total, 600 de los audios del corpus corresponden a hablantes hispanohablantes. Esto corresponde a aproximadamente 1,132 oraciones con una duración promedio de 3.6 segundos por oración y alrededor de nueve palabras por cada una. No se proporciona información específica sobre la edad ni origen de los hablantes, pero se puede inferir que se trata de adultos al escuchar las grabaciones [39].

EpaDB

Se compone de 3200 enunciados provenientes de 50 hablantes de español de Argentina, divididos en partes iguales entre 25 mujeres y 25 hombres. Las grabaciones consisten en 64 frases cortas en inglés diseñadas específicamente para proporcionar una muestra equilibrada de los fonemas del idioma, con un enfoque particular en aquellos que suelen ser pronunciados de manera incorrecta por hablantes nativos de español de la región.

Cuenta con transcripción a nivel de fonema utilizando símbolos *ARPAbet* y un conjunto adicional de extensiones necesarias para representar variantes no nativas encontradas. Contiene una alineación hecha a mano y calificaciones correspondientes a la percepción global del grado de natividad en una escala de 1 a 5 [40]. Dicha información se presenta en archivos *TextGrid*.

Esta base de datos se considera de alta relevancia ya que el acento de los hablantes es similar al del público objetivo, español Rioplatense. Es por ello que se utilizará tanto para la validación de métodos de evaluación de pronunciación como para la selección del transcriptor de voz a texto.

speechocean762

Base de datos diseñada específicamente para la evaluación de la pronunciación, compuesta por 5000 enunciados en inglés de 250 hablantes no nativos, con la particularidad de que la mitad de los hablantes son niños [28].

Esta base de datos ha sido anotada a nivel de frase, palabra y fonema por cinco expertos. Las anotaciones fueron realizadas bajo un conjunto de métricas definidas, evaluando la precisión de la pronunciación, fluidez, completitud y prosodia en una escala del 0 al 10. Los hablantes son nativos de habla mandarín, lo cual le quita utilidad en nuestro caso, ya que los patrones de error en la pronunciación que pueden tener hablantes nativos de español pueden diferir significativamente con respecto a los errores que estos hablantes puedan presentar, debido a las diferencias en los sistemas fonológicos de cada lengua. De

todas formas, es un recurso interesante en el que se entrenan varios modelos de evaluación de pronunciación como *GOPT* o *MultiPA*.

Base de datos privada

Ceibal

Se cuenta con un conjunto pequeño de audios provenientes de pruebas realizadas por el programa *Ceibal en Inglés* con grabaciones de 23 estudiantes. Cada estudiante aporta 14 audios de duraciones variadas no mayores a 30 segundos, sumando 322 archivos.

Estas grabaciones son respuestas a preguntas específicas de la prueba y, aunque cada estudiante proporciona múltiples audios, la calificación se asigna a nivel global por estudiante, no por audio individual. Los evaluadores otorgan una puntuación holística en una escala del 0 al 9, que surge a partir de la suma de puntajes también holísticos del 0 al 3 en tres categorías: *Task Fulfilment*, *Language (Range & Accuracy)* y *Delivery (Intelligibility & Flow of Speech)*. En esta sección, la categoría de interés es *Delivery*, ya que refleja la evaluación de características acústicas, teniendo en cuenta la inteligibilidad y fluidez de la respuesta.

Este conjunto de datos se caracteriza por incluir habla espontánea, resultado directo de las preguntas formuladas a los estudiantes, quienes tienen la libertad de responder de manera abierta, reflejando así su capacidad de expresión oral real.

No se cuenta con etiquetas más allá de la evaluación antes mencionada. La ausencia de etiquetas detalladas para cada audio, la poca cantidad de audios y el hecho de que no todos sean útiles, ya que algunos no contienen voz o están en español, limita la capacidad de utilizar estos datos para entrenar o ajustar modelos.

De todas formas, esta base es de gran utilidad para construir un conjunto de validación y probar herramientas, ya que contiene audios de la población de interés.

Selección de bases de datos a utilizar para la evaluación

En base a los objetivos del estudio y al perfil del público objetivo, se seleccionaron dos bases de datos para la evaluación: EpaDB, compuesta por hablantes de inglés cuya lengua materna es el español rioplatense, al igual que en nuestro caso, y la base de datos de Ceibal, que proviene directamente del contexto local y contiene grabaciones reales del público objetivo. Por otro lado, si bien L2-ARCTIC incluye cuatro hablantes con lengua materna español, se trata de adultos provenientes de España, por lo que su representatividad para este trabajo es limitada. Por tanto, dicha base no fue utilizada para la evaluación de los métodos implementados.

3.2.2. Disponibilidad de Código con Capacidad de Inferencia

Un aspecto clave en la selección de los métodos es la *disponibilidad de código*. Si bien existen numerosas publicaciones sobre evaluación automática de pronunciación, una de las principales dificultades encontradas es que la mayoría de estos trabajos no incluyen código asociado ni explican claramente cómo implementar los métodos descritos. Además, muchos estudios presentan resultados basados en una base de datos específica, controlada y no pública, lo que limita aún más su aplicabilidad en otros contextos, ya que estos métodos suelen estar diseñados exclusivamente para fines de investigación y no para ser implementados de manera práctica.

Además de la disponibilidad de código, es fundamental que los métodos seleccionados ofrezcan la posibilidad de realizar inferencias sobre audios propios. Esto implica que, aunque un método cuente con una implementación pública, debe ser adaptable y flexible para trabajar con datos diferentes a los de las bases de datos originales utilizadas en las investigaciones. En muchos casos, la inferencia sobre datos nuevos no es un proceso medianamente directo o realizable, lo que limita la aplicabilidad.

Por tanto, los métodos seleccionados deben tener código accesible y permitir realizar inferencias sobre conjuntos de datos diferentes a los utilizados en los estudios originales. En la siguiente sección se presenta y profundiza en los métodos seleccionados.

3.2.3. Implementación de Métodos

Se implementaron dos enfoques principales para la evaluación de pronunciación: el desarrollo de una *línea base* y la evaluación del modelo *MultiPA* en nuestro contexto.

El primer enfoque consiste en construir una *línea base* que, a partir de un audio y un texto de referencia, permita comparar la secuencia de fonemas detectados en el audio con los fonemas de la pronunciación canónica del texto. A pesar de las limitaciones de ese tipo de enfoques se decide implementarlo debido a su valor para la adaptación y comprensión del proceso de evaluación.

El segundo enfoque es la implementación del modelo “**MultiPA**: A Multi-task Speech Pronunciation Assessment System for Closed and Open Response Scenario”, un sistema avanzado que permite la evaluación de diversos aspectos de la pronunciación en escenarios de respuesta cerrada y abierta.

A continuación, se explican en detalle ambos enfoques en las subsecciones correspondientes.

Línea Base

Se centra en la comparación de secuencias de fonemas. Este enfoque se enmarca dentro de los **métodos de reconocimiento canónico**, ya que su objetivo principal es determinar si la pronunciación del hablante se adhiere a una secuencia estándar o canónica de fonemas, derivada del texto de referencia. Por esto último es que además entra dentro de la clasificación de **métodos de evaluación de habla leída**.

En la figura 3.4 se presenta el diagrama de la arquitectura de la línea base.

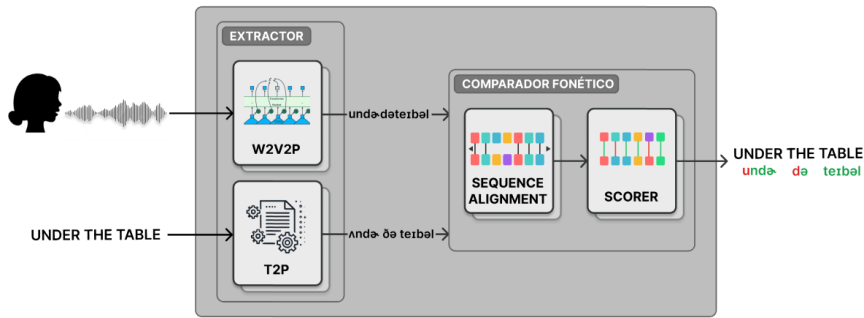


Figura 3.4: Arquitectura de la línea base.

El proceso inicia con la extracción de fonemas a partir de la señal de audio utilizando un modelo preentrenado basado en *Wav2Vec2*, llamado *Wav2Vec2Phonemes* (W2V2P en la figura 3.4), ajustado para la extracción de fonemas [41]. Resaltar que la secuencia obtenida por este modelo no se presenta alineada por palabra.

Paralelamente, la secuencia de fonemas esperados se obtiene a partir de la transcripción del texto utilizando el módulo *TextToPhonemesExtractor*, que emplea *Phonemizer* con *eSpeak* como backend para realizar la conversión [42,43]. Esta secuencia sí se obtiene por palabra. En la figura 3.4, este módulo es llamado T2P.

Una vez obtenidas ambas secuencias fonéticas, el sistema procede a alinear la secuencia de fonemas obtenidos a partir del audio, minimizando el error respecto a la secuencia de fonemas por palabra. Esto permite obtener, por cada palabra, una secuencia de fonemas esperados y una secuencia de fonemas escuchados.

Luego, se comparan los fonemas detectados con la pronunciación canónica del texto de referencia, representado por *Sequence Alignment* en la 3.4. Para esta comparación, se utilizan tres métodos: dos variantes de la *distancia de Levenshtein* y la técnica *Dynamic Time Warping (DTW)*. Entre los métodos asociados a Levenshtein se encuentran: la versión estándar, que mide las diferencias entre las secuencias de fonemas en términos de inserciones, eliminaciones y sustituciones, y una versión ponderada, en la que los costos de estas operaciones se ajustan según la similitud acústica entre los fonemas. Esta similitud acústica

se refiere a qué tan parecidos son dos fonemas en el sonido que producen. Por ejemplo, los fonemas /p/ y /b/ suenan similares porque ambos implican una obstrucción en los labios y una breve liberación de aire, aunque se diferencian en que /b/ es un sonido sonoro (vibran las cuerdas vocales) y /p/ no lo es.

El módulo *PhonemeScorer* (**Scorer** en la figura 3.4) se encarga de gestionar esta ponderación, asignando menores costos a las sustituciones entre fonemas que son acústicamente similares. Esto permite que el sistema sea más tolerante a variaciones fonéticas leves. De esta forma, no solo se consideran las discrepancias literales entre los fonemas producidos y los esperados, sino también las similitudes fonéticas entre ellos.

El resultado de estas comparaciones es un conjunto de puntuaciones que cuantifican la precisión de la pronunciación del hablante con respecto a la secuencia canónica de fonemas. Estas puntuaciones se calculan tanto a nivel global, considerando la totalidad de la secuencia, como a nivel de palabra.

A pesar de las limitaciones inherentes a este enfoque, como su dependencia de la comparación fonética literal y la dificultad para manejar las variaciones del habla espontánea, se decide incluir esta línea base por varias razones. En primer lugar, este método proporciona un punto de partida claro y comprensible para la evaluación automática de la pronunciación, permitiendo una adaptación gradual hacia métodos más sofisticados. Además, aunque este enfoque está diseñado para evaluar la pronunciación en el contexto de **habla leída** (una modalidad que no se evalúa actualmente en la prueba de *Ceibal en Inglés*) es importante destacar que la mayoría de los sistemas de evaluación automática de la pronunciación se centran precisamente en el habla leída.

MultiPA

Dentro del enfoque de modelos basados en aprendizaje auto-supervisado (*Self-Supervised Learning - SSL*), presentado en la sección 3.1.1, *MultiPA* (*Multi-task Pronunciation Assessment*) es un modelo diseñado para evaluar simultáneamente múltiples aspectos de la pronunciación. Este modelo proporciona evaluación de precisión, fluidez y prosodia a nivel de oración, así como evaluación de precisión a nivel de palabra en escenarios de respuestas abiertas y cerradas [34]. En lugar de entrenar un modelo separado para cada una de estas características, las evalúa de manera conjunta en una sola arquitectura, lo que se conoce como aprendizaje multi-tarea (*multi-task learning*).

El modelo utiliza *HuBERT: Hidden-Unit BERT*, un sistema preentrenado de aprendizaje auto-supervisado [44], ajustado específicamente para realizar tareas multi-tarea en la evaluación de la pronunciación. MultiPA también emplea el sistema de reconocimiento automático del habla (ASR) *Whisper*, para generar transcripciones, y la herramienta de alineación fonética *Charsi* [45], que se encarga de alinear las palabras con la señal de habla [34]. El entrenamiento del modelo se llevó a cabo utilizando el conjunto de datos *speechocean762*. Esto permitió ajustar las capas de *Transformer Encoders* y las capas convolucionales y lineales.

A continuación, se presenta en la Figura 3.5 el esquema de la arquitectura de *MultiPA*. El proceso comienza con la extracción de características, donde se utilizan dos modelos ASR de *Whisper*. El modelo *base.en*, de aproximadamente 74 millones de parámetros, es más liviano y rápido, y se emplea para generar la transcripción automática del audio producido por el estudiante. Por su parte, el modelo *medium.en*, con cerca de 769 millones de parámetros, ofrece mayor precisión y se utiliza para obtener la transcripción del enunciado objetivo durante el entrenamiento. En el caso *close response* no se utiliza el modelo Medium ya que se cuenta con el texto de referencia y es ésta la transcripción objetivo.

Posteriormente, se utiliza *Charsiu* para alinear las palabras entre las transcripciones y las señales de audio. Después, *RoBERTa* [46] se emplea para generar representaciones semánticas de las transcripciones, lo que proporciona información adicional sobre el significado de las palabras y ayuda a mejorar la precisión de las transcripciones y alineaciones obtenidas.

La estructura principal del modelo se basa en *HuBERT*, que extrae representaciones latentes del habla. Estas representaciones capturan tanto las transiciones entre fonemas y la información prosódica, como la entonación y el ritmo. Luego, estas representaciones son procesadas mediante capas de *Transformer Encoders*, que fusionan las características obtenidas, utilizando promediado (*average pooling*) para alinear las características de nivel de palabra y fonema. Finalmente, las evaluaciones se realizan mediante capas convolucionales y lineales para obtener las métricas de precisión, fluidez y prosodia a nivel de oración y palabra.

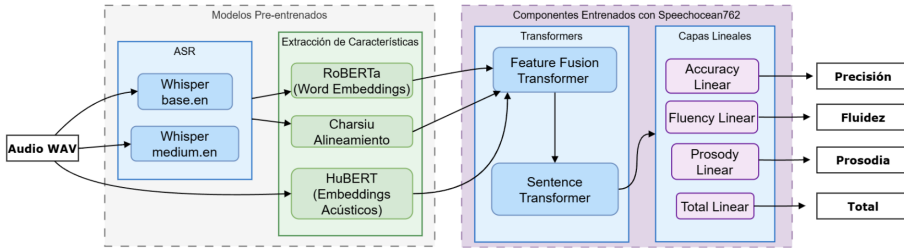


Figura 3.5: Esquema general de la arquitectura de MultiPA.

El diagrama muestra las distintas fases del modelo, desde la extracción de características utilizando *Whisper* y *Charsiu*, hasta el uso de *HuBERT* como modelo acústico, para finalmente realizar evaluaciones a nivel de oración y palabra.

Una de las características fundamentales de *MultiPA* es su capacidad para trabajar en escenarios de respuestas abiertas, donde la transcripción del habla no está disponible de antemano. Esto lo hace adecuado para el contexto de la prueba de *Ceibal en Inglés*, ya que las respuestas son espontáneas a partir de preguntas definidas.

Al procesar un audio con *MultiPA*, el modelo genera varias salidas en formato *JSON* que incluyen tanto evaluaciones globales como a nivel de palabra.

Entre las métricas a nivel de oración se encuentran *uttr_acc* (precisión de la pronunciación), *uttr_fluency* (fluidez), *uttr_prosodic* (prosodia) y *uttr_total* (puntaje total de la oración). A nivel de palabra, se obtienen valores de *word_acc* (precisión de la pronunciación de cada palabra), *word_stress* (evaluación de la correcta colocación del acento prosódico en las palabras) y *word_total* (puntaje global de cada palabra). Cabe aclarar que las salidas asociadas a puntajes globales, como *uttr_total* y *word_total* surgen a partir de predecir etiquetas también globales de la base *speechocean762*.

Además, se proporcionan las transcripciones de la oración en dos formatos: en una respuesta cerrada, *transcript_W* corresponde al texto dado, mientras que en respuesta abierta se utiliza el modelo Whisper medium. Por otro lado, *transcript_S* siempre es generado por el modelo Whisper base.

Los autores reportan que los experimentos realizados con *MultiPA* en diferentes conjuntos de datos, tanto en aquellos similares a los utilizados durante el entrenamiento como en otros con características distintas, han demostrado que el modelo logra un rendimiento superior a otros enfoques, como *GOPT*.

En la siguiente sección, se hará inferencia con *MultiPA* utilizando bases de datos de hablantes hispanohablantes del Río de la Plata, ya que este es el público objetivo de la evaluación. *MultiPA* es capaz de evaluar el habla en distintos niveles, y para esta validación se optó por realizar una evaluación a nivel de frase, ya que es la más alineada con la metodología actual de Ceibal.

3.3. Evaluación de Pronunciación

3.3.1. Evaluación de la Línea Base

En esta sección se presentan los resultados al evaluar la línea base en la base de datos *EpaDB*. La elección de esta base se debe a que, como se mencionó en la sección 3.2.1 corresponde a habla leída y presenta etiquetas a nivel de fonema y global.

Para ilustrar el análisis, en primer lugar se presenta el resultado de la inferencia en un audio de dicha base. En particular, se elige el audio número 52 correspondiente al speaker 14, cuya frase es “Let’s play soccer”.

En la Figura 3.6, se presentan los resultados de la línea base para el audio seleccionado de *EpaDB*, desglosados por palabra. Se incluyen las métricas de distancia de Levenshtein, Levenshtein ponderado y DTW. Los puntajes normalizados muestran estas métricas en una escala común (de 0 a 1), para facilitar la comparación entre palabras. Así, un valor cercano a 1 indica una coincidencia mayor con la pronunciación esperada, mientras que valores más bajos reflejan mayores diferencias.

Por ejemplo, para la primera palabra “let’s”, la distancia de Levenshtein es 3 porque se necesitan tres operaciones para transformar la secuencia de fonemas esperada *lets* en la secuencia predicha *les*, y la penalización uniforme de Levenshtein asigna un costo de 1 a cada operación. Las operaciones específicas

Per Word Scores

Word	Expected	Predicted	Levenshtein	Weighted Levenshtein	DTW	Normalized Scores
let's	lets	les	3.00	0.05	2.00	L:0.25, WL:0.99, DTW:0.50
play	pler	pler	0.00	0.00	0.00	L:1.00, WL:1.00, DTW:1.00
soccer	sa:ke	soker	4.00	1.14	4.00	L:0.20, WL:0.77, DTW:0.20

Figura 3.6: Resultado por palabra de la línea base en un audio de EpaDB.

son las siguientes:

- Primero, se realiza una **sustitución** del fonema ϵ por el fonema e.
- Luego, se realiza una **sustitución** del fonema t por el fonema s.
- Finalmente, se realiza una **eliminación** del fonema s de la secuencia esperada.

El *Weighted Levenshtein* da un valor de 0.05 para esta palabra, significativamente menor que Levenshtein general. Esto puede deberse fundamentalmente a que la diferencia fonética entre ϵ y e es pequeña, por lo que la sustitución tiene un costo reducido.

Por otro lado, DTW da un valor de 2.00 para esta misma palabra. Este método busca la alineación óptima entre ambas secuencias, permitiendo “estiramientos” o “compresiones” de los fonemas para minimizar el costo de la comparación. Así, permite encontrar la mejor correspondencia entre los fonemas, lo que resulta en una penalización menor en comparación con Levenshtein.

Para la segunda palabra “play”, tanto Levenshtein como DTW dan una puntuación de 0.00, lo que indica que la secuencia de fonemas esperada y la predicha son idénticas: pler. En este caso, ambas métricas coinciden en que no hay penalización.

Para la tercera palabra “soccer”, la distancia de Levenshtein es 4, ya que se requieren cuatro operaciones para alinear las secuencias. DTW también da un valor de 4.00, lo que indica que ambas métricas coinciden en este caso en la penalización. En este caso, DTW no puede encontrar una alineación menos costosa que la obtenida por Levenshtein.

Para la evaluación de este método se recurre a las etiquetas a nivel de fonema presentes en la base *EpaDB*. Para cada audio, se tiene un archivo .csv llamado “labels” que contiene los fonemas en formato *ARPAbet* de la frase, junto con un símbolo de más (+) o menos (-) que indican si el fonema esperado coincide con el realmente dicho. En la tabla 3.1 se presenta el etiquetado a nivel de fonema para la frase “Let’s play soccer”. Es necesario notar que esta etiqueta no tiene en cuenta la distancia entre fonemas. Dos pares de fonemas distintos llevarán la misma etiqueta (-) aunque en un caso sean fonéticamente más similares.

Para cada frase se tiene entonces un conjunto de símbolos de más y menos que contienen la información de coincidencia entre fonemas. La cantidad de discrepancias es el valor que tomaremos en cuenta para la evaluación.

Fonema esperado	Fonema dicho	Coincidencia
L	L	+
EH	EH	+
T	T	+
S	S	+
P	P	+
L	L	+
EY	EY	+
S	S	+
AO	AO	+
K	Kh	-
ER	EH+DX	-

Tabla 3.1: Tabla de fonemas comparados para una frase del speaker 14 de la base EpaDB.

Para evaluar tanto este método como los siguientes a lo largo del trabajo, se define un subconjunto de la base *EpaDB* el cual consiste en 1332 audios de una muestra aleatoria de 21 hablantes. Este subconjunto a partir de ahora se llamará *EpaDB subset*.

Dicha base de datos cuenta con alineación a nivel de palabra, lo que permite asociar cada fonema a su palabra correspondiente y, de esta forma, calcular una puntuación por palabra.

Para la evaluación a nivel de palabra, se construyen dos vectores con longitud equivalente a la suma de la cantidad de palabras de todos los hablantes de esta muestra. Los vectores contienen:

- **Vector de línea base:** el puntaje DTW por cada palabra.
- **Vector de referencia:** para cada palabra, la cantidad de etiquetas (-) correspondiente, de la base EpaDB.

La elección de DTW como métrica para el cálculo de la correlación en este estudio se justifica por las características de las etiquetas disponibles en la base de datos *EpaDB*. DTW permite comparar dos secuencias —en este caso, secuencias de fonemas— buscando una alineación óptima que minimice la distancia total entre ellas, incluso si hay desplazamientos en el orden de los elementos. A diferencia de Levenshtein, que penaliza de forma estricta cada operación de inserción, eliminación o sustitución, DTW permite una correspondencia más flexible entre fonemas, lo que evita penalizaciones excesivas cuando la secuencia reconocida está ligeramente desfasada respecto a la esperada. Esto hace que los puntajes de Levenshtein tiendan a ser mayores que los reflejados en las etiquetas de *EpaDB*, generando una discrepancia con los errores esperados. Además, el uso de DTW resultó en una correlación más alta en comparación con Levenshtein, lo que respalda su idoneidad frente a la otra métrica para este análisis.

En la tabla 3.2 se presentan los resultados de las correlaciones de Pearson y Spearman obtenidas tanto a nivel de palabra como de frase, junto con sus respectivos valores p.

Nivel	Tipo	Correlación	Valor p
Palabra	Pearson	0.3608	$1,18 \times 10^{-210}$
	Spearman	0.2527	$5,26 \times 10^{-101}$
Frase	Pearson	0.3768	$1,30 \times 10^{-45}$
	Spearman	0.2791	$2,99 \times 10^{-25}$

Tabla 3.2: Correlaciones de Pearson y Spearman y valores p para los niveles de palabra y frase en la evaluación de coincidencia fonética.

Ambos coeficientes van de -1 a 1, donde un valor de 1 indica una relación positiva perfecta, -1 indica una relación negativa perfecta y 0 indica que no hay relación entre las variables.

La categorización de los valores absolutos de los coeficientes de correlación es la siguiente [47]:

- **0.7 < 1.0:** Correlación muy fuerte.
- **0.5 < 0.7:** Correlación fuerte.
- **0.3 < 0.5:** Correlación moderada.
- **0.1 < 0.3:** Correlación débil.
- **0.0 < 0.1:** Sin correlación.

Adicionalmente, ambos coeficientes se acompañan de un valor p, que indica la significancia estadística del coeficiente de correlación obtenido. Un valor p bajo, generalmente menor a 0.05, sugiere que la correlación observada es significativa y no se debe al azar.

El cálculo de ambos coeficientes de correlación, Pearson y Spearman, se justifica por la necesidad de evaluar diferentes tipos de relaciones entre las puntuaciones DTW y las etiquetas de la base *EpaDB*. Pearson mide correlaciones lineales, mientras que Spearman evalúa correlaciones monotónicas, capturando relaciones que pueden no ser lineales pero que mantienen un orden consistente entre los datos. Esta dualidad es particularmente relevante dado que no se puede asumir a priori que la relación entre las puntuaciones y las etiquetas sea estrictamente lineal. Además, Spearman, al calcularse en función de rangos, es menos sensible a valores atípicos, complementando la sensibilidad de Pearson a relaciones lineales.

Los resultados presentados en la Tabla 3.2 muestran correlaciones de Pearson moderadas y correlaciones de Spearman débiles, siendo todas estadísticamente significativas tanto a nivel de palabra como de frase. La correlación de Pearson es ligeramente mayor en ambos niveles, lo que indica una relación principalmente lineal entre los puntajes de distancia fonética calculados por el modelo y las

discrepancias identificadas en las etiquetas de la base *EpaDB*. Esto sugiere que, a medida que aumenta la distancia fonética según el modelo, también aumenta proporcionalmente la cantidad de discrepancias reales.

Por otro lado, la base *EpaDB* también proporciona un puntaje global correspondiente al grado de natividad detectado en cada audio. Se calcularon también correlaciones respecto a este puntaje.

Para ello, se tomaron los vectores anteriores con puntajes por frase y se normalizaron respecto a la cantidad de fonemas de cada frase. Esta normalización es pertinente porque se infiere que el puntaje global indirectamente considera la longitud de la frase.

Procedemos entonces a calcular las correlaciones. En primer lugar, para comparar etiquetas dentro de la misma frase proporcionadas por la base de datos, se calculó la correlación entre el vector que contiene la proporción de fonemas no correspondidos (Vector etiquetas fonéticas) y la etiqueta de grado de natividad.

Luego, se calculó la correlación entre el vector de puntajes normalizados predichos por la línea base (Vector línea base) y el grado de natividad.

Los resultados de estas correlaciones se presentan en la Tabla 3.3.

Variable	Tipo	Correlación	Valor p
Vector etiquetas fonéticas	Pearson	0.5904	$7,38 \times 10^{-126}$
	Spearman	0.6042	$3,05 \times 10^{-133}$
Vector línea base	Pearson	0.2277	$6,48 \times 10^{-17}$
	Spearman	0.2052	$5,77 \times 10^{-14}$

Tabla 3.3: Correlaciones de Pearson y Spearman entre los puntajes normalizados y el grado de natividad.

Como se observa en la tabla 3.3, la correlación de Pearson entre el vector de etiqueta fonética de *EpaDB* y el grado de natividad es de 0.5904, mientras que la correlación de Spearman es de 0.6042. La similitud entre los valores de Pearson y Spearman sugiere que la relación entre las etiquetas fonéticas y el grado de natividad es predominantemente monotónica, sin grandes desviaciones de linealidad. Estos valores indican una concordancia de magnitud fuerte. Sin embargo, el hecho de que la correlación no sea más alta podría indicar que estas etiquetas, basadas en discrepancias fonéticas, capturan aspectos de la pronunciación que no reflejan por completo el grado de natividad. Esto podría deberse a que se enfocan en diferencias fonéticas particulares, sin considerar otros aspectos prosódicos o semánticos que también influyen en la percepción de natividad.

Por otro lado, al comparar el vector resultante de la línea base, las correlaciones de Pearson y Spearman son de 0.2277 y 0.2052 respectivamente. Estas correlaciones positivas, aunque de baja magnitud, son significativas debido al valor p asociado, y reflejan una relación débil con el grado de natividad.

Estas limitaciones pueden deberse a varios factores. En primer lugar, la extracción de fonemas a partir de la señal de audio utilizando el modelo *Wav2Vec2-Phonemes*, aunque efectiva, puede no ser suficientemente robusta frente a va-

riaciones en la pronunciación o acentos diferentes. Además, la representación discreta de los fonemas puede no reflejar plenamente las transiciones suaves y coarticulaciones presentes en el habla natural. Las métricas utilizadas, como Levenshtein y DTW pueden ser sensibles a errores de alineación y no capturar completamente las similitudes fonéticas perceptuales.

Esta baja correlación resulta coherente con el hecho de que las propias etiquetas fonéticas de EpaDB, diseñadas para capturar discrepancias entre fonemas, muestran solo una correlación moderada con el grado de natividad. Es decir, dado que ni siquiera las etiquetas internas de la base logran una correlación alta con el grado de natividad, es comprensible que el vector de la línea base, también exhiba una correlación limitada. Estos resultados sugieren que, aunque existe una relación entre los puntajes normalizados y el grado de natividad, las etiquetas fonéticas y la forma en que están medidas por sí solas no capturan todos los aspectos de la pronunciación que contribuyen a la percepción de natividad.

De todas formas, cabe aclarar que la razón por la cual se está comparando con grado de natividad es debido a que es una de las etiquetas disponibles para poder comparar que pueda captar aspectos de *Delivery* en el habla.

Las bajas correlaciones halladas en este análisis reflejan que es necesario explorar métodos más sofisticados para la evaluación de la pronunciación. Incorporar modelos que integren representaciones continuas de las características acústicas y contextuales.

En la siguiente sección, se presenta **MultiPA**, una técnica más avanzada para la evaluación de la pronunciación, que aborda algunas de estas limitaciones al considerar de manera más integral las características fonéticas y acústicas del habla.

3.3.2. Evaluación de MultiPA

En esta sección se presentan los resultados de la inferencia del método MultiPA aplicados a audios de las bases de datos *EpaDB* y Ceibal. Estas bases son fundamentales para evaluar cómo se comporta el modelo con hablantes que tienen acentos similares o representan al público objetivo del estudio. Dado que MultiPA fue entrenado originalmente con hablantes cuya lengua materna es mandarín y no está ajustado al acento rioplatense ni al español en general, es necesario comprobar su efectividad en este contexto.

Es importante señalar que *EpaDB* y Ceibal son las únicas bases de datos disponibles con hablantes nativos de español (L1) y con etiquetas globales, lo que justifica su uso en esta investigación.

El análisis comparativo de los resultados en ambas bases tiene como objetivo examinar la robustez de MultiPA en este escenario lingüístico, que difiere significativamente de los datos con los que fue entrenado.

El proceso de evaluación seguirá los siguientes pasos:

1. Selección de audios a utilizar y pre-procesamiento en caso necesario.
2. Inferencia en MultiPA.

3. En caso necesario, obtención de métricas agregadas a partir de los resultados de la inferencia.
4. Comparación de resultados con evaluaciones de expertos humanos.
5. Análisis de casos específicos.

Evaluación sobre EpaDB

En este apartado se realiza el proceso de evaluación sobre la base EpaDB.

Se utilizará el puntaje global proporcionado por *EpaDB* para cada audio, el cual refleja el nivel de similitud con el habla de un nativo percibido por los evaluadores. Este criterio difiere del de inteligibilidad, aunque parcialmente lo incluye; se espera que una persona con un alto nivel de similitud con un nativo también sea inteligible. Por el contrario, es posible que una persona sea inteligible sin necesariamente hablar de manera similar a un nativo.

El subconjunto de audios con el cual se realiza inferencia es *EpaDB subset*, definido en la sección 3.3.1. Esta vez, se utilizará la etiqueta global de dicha base, correspondiente al grado de natividad en cada audio.

En la figura 3.7 se muestra la distribución de los puntajes asignados a cada audio.

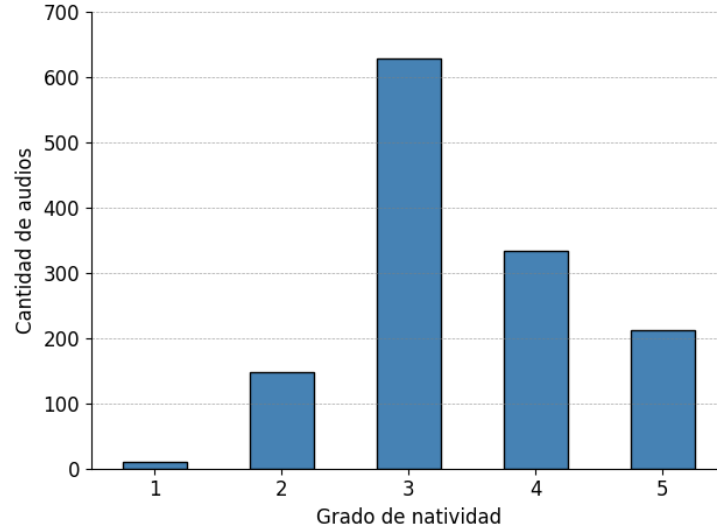


Figura 3.7: Histograma de puntaje global en las muestras del subconjunto de la base EpaDB.

Se puede observar que existe un claro desbalance entre las clases, siendo el nivel medio el más representado, con 628 muestras, mientras que 11 audios son considerados de baja similitud con el habla de un nativo, correspondientes a un puntaje de 1.

No fue necesario realizar ningún preprocesamiento en esta base de datos, ya que todos contienen palabras en inglés y no presentan nivel de ruido que pueda afectar el análisis.

Se procede a inferir en MultiPA. En este caso, se quiere evaluar a nivel de frase, por lo tanto, las salidas de interés son: *uttr_acc*, *uttr_fluency*, *uttr_prosodic* y *uttr_total*.

Análisis cuantitativo

El objetivo de esta sección es explorar la relación entre las características evaluadas por MultiPA y las etiquetas de evaluación global de EpaDB. Para ello, se llevarán a cabo una serie de análisis cuantitativos con el fin de examinar la distribución de las características individuales, las interrelaciones entre ellas y la posible relación con las etiquetas.

Las etiquetas globales en *EpaDB* y las puntuaciones del modelo se encuentran en escalas que no solo difieren en su rango, sino también en su enfoque de evaluación. Las etiquetas van de 1 a 5 y reflejan una evaluación holística realizada por evaluadores humanos, mientras que las puntuaciones del modelo, que van de 0 a 10, resultan de un análisis detallado de aspectos específicos como la precisión, fluidez y prosodia. Debido a estas diferencias en cómo se interpretan las escalas, no es posible realizar una comparación directa entre los valores. Es necesario investigar si existe alguna relación entre ambas evaluaciones y si el modelo puede diferenciar de manera efectiva entre los distintos grados de natividad a través de sus puntuaciones. Este análisis permitirá evaluar si las puntuaciones reflejan adecuadamente las diferencias percibidas por los evaluadores humanos y si son lo suficientemente consistentes para discriminar entre los niveles de natividad.

Análisis descriptivo

En esta sección, se realiza un análisis descriptivo de las características evaluadas por el modelo en relación con el grado de natividad de los hablantes. Este análisis incluye la evaluación de las características de Precisión, Fluidez, Prosodia y Puntaje Total. Se utilizan estadísticas descriptivas, como la media, mediana, desviación estándar y rangos (mínimo y máximo) para cada característica, agrupadas según el grado de natividad, que varía en una escala de 1 a 5. Los resultados detallados de este análisis se presentan en el anexo [A.1](#) en formato de tablas.

El propósito es determinar si el modelo es capaz de diferenciar de manera efectiva entre los distintos niveles de natividad. Esto se evalúa observando si las puntuaciones asignadas a cada característica aumentan de forma coherente con el grado de natividad. Además, permite identificar posibles anomalías, como una alta variabilidad o la presencia de valores atípicos, que podrían indicar inconsistencias en la evaluación del modelo.

Para complementar el análisis, se generaron diagramas de caja para cada característica, agrupados según el grado de natividad. Proporcionan una repre-

sentación visual de la mediana, los cuartiles y la dispersión de las puntuaciones, así como la presencia de valores atípicos. Se presentan en la figura 3.8.

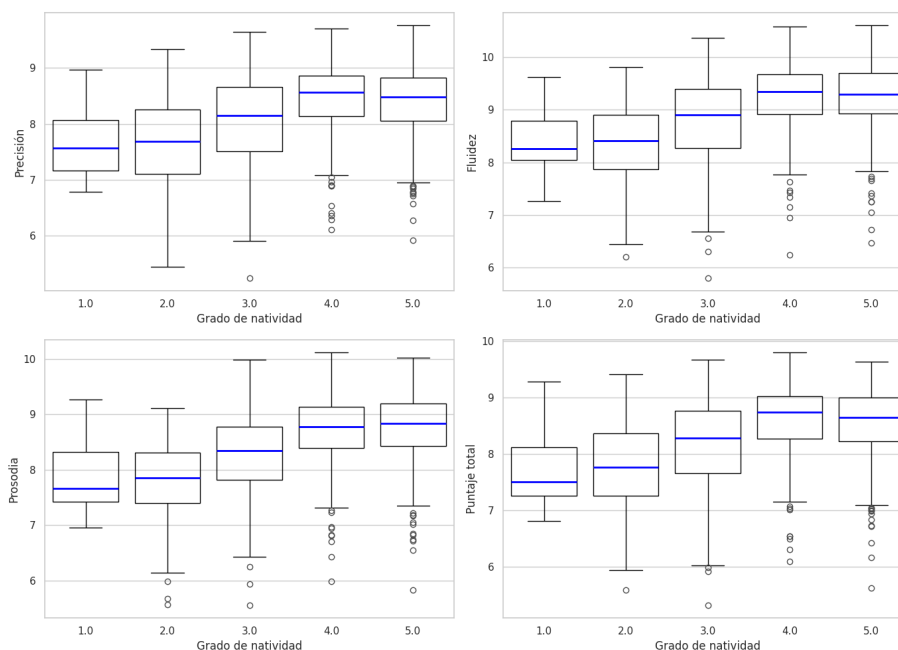


Figura 3.8: Distribución de Precisión, Fluidez, Prosodia y Total en función de las etiquetas en EpaDB.

Los diagramas de caja muestran que, si bien las medianas de las puntuaciones tienden a aumentar ligeramente con el grado de natividad, este cambio no parece ser suficientemente significativo cuando se considera el rango completo de puntuación posible. Las puntuaciones se concentran en un rango bastante estrecho, que va aproximadamente desde 5.5 hasta 10, dejando una parte considerable del rango, especialmente los valores más bajos, sin utilizar. Esto sugiere que el modelo tiene una capacidad limitada para discriminar entre niveles de natividad, ya que las diferencias en las puntuaciones son pequeñas en comparación con el rango total.

En cuanto a la variabilidad, se observa que esta es moderada en la mayoría de los grupos de natividad, con algunos valores atípicos presentes especialmente en los niveles más altos (Grado 4 y 5). En la siguiente sección se estudian dichos casos atípicos.

Precisión: Los resultados muestran que la media de `uttr_acc` aumenta de 7.71 en grado de natividad 1 a 8.48 en grado 4. Aunque este incremento sugiere que el modelo intenta discriminar entre diferentes grados de natividad, el cambio en las puntuaciones es relativamente pequeño considerando el rango completo de puntuación (0 a 10). La disminución observada en puntaje 5, donde

la media baja a 8.38, podría indicar que el modelo enfrenta dificultades para diferenciar entre hablantes con natividad muy alta. Sin embargo, dado el rango limitado en el que operan las puntuaciones, estas variaciones podrían no ser lo suficientemente significativas para permitir una clasificación clara entre los distintos grados de natividad.

Fluidez: La media de `uttr_fluency` muestra un aumento de 8.35 en grado de natividad 1 a 9.25 en grado 4, estabilizándose en grado 5. Este comportamiento sugiere que el modelo es relativamente más eficaz en captar variaciones en fluidez a medida que aumenta la natividad, pero, al igual que en Precisión, el rango de puntuación utilizado es limitado, lo que implica que las diferencias observadas pueden no ser lo suficientemente significativas en el contexto del rango total disponible. La menor variabilidad en los niveles superiores podría indicar una evaluación más consistente por parte del modelo en estos grupos, pero la diferencia global sigue siendo sutil.

Prosodia: La media de `uttr_prosodic` aumenta de manera constante desde 7.84 en grado 1 hasta 8.72 en los grados 4 y 5. No obstante, como en las características anteriores, el aumento es relativamente pequeño cuando se considera el rango completo de puntuación. La desviación estándar constante a través de los niveles sugiere que el modelo mantiene una evaluación consistente, aunque nuevamente, las diferencias son moderadas y pueden no ser suficientes para una clara clasificación de los grados de natividad.

Total: El puntaje total muestra una tendencia de incremento general, alcanzando un máximo en grado de natividad 4 (8.61) seguido de una ligera disminución en 5 (8.52). Luego, esta característica sigue la misma tendencia que las anteriores.

Análisis de correlación

Luego de realizar un análisis descriptivo básico, se procede a estudiar si existe correlación entre las etiquetas a nivel de frase de *EpaDB* y la salida del modelo. Para ello, se calculan los coeficientes de correlación de Pearson y Spearman, cuyos resultados se presentan en la tabla 3.4.

Métrica a nivel de frase	Pearson		Spearman	
	Coeficiente	Valor p	Coeficiente	Valor p
Precisión	0.31	$3,89 \times 10^{-31}$	0.32	$2,83 \times 10^{-33}$
Fluidez	0.35	$1,31 \times 10^{-40}$	0.38	$4,04 \times 10^{-46}$
Prosodia	0.38	$3,54 \times 10^{-48}$	0.41	$3,04 \times 10^{-56}$
Total	0.32	$6,91 \times 10^{-34}$	0.34	$3,22 \times 10^{-38}$

Tabla 3.4: Coeficientes de correlación y valores p de Pearson y Spearman para las características obtenidas en la inferencia de MultiPA con un subconjunto de la base EpaDB.

El coeficiente de correlación de Pearson se utiliza debido a su amplia aceptación en la comparación de métodos de pronunciación; sin embargo, no es robusto

frente a valores atípicos y asume tanto una relación lineal entre las variables como una distribución normal de los residuos, supuestos que en este contexto no se pueden garantizar.

Dado que las predicciones del modelo son continuas mientras que las etiquetas son discretas, el coeficiente de correlación de Spearman se considera más adecuado para este análisis. Mide relaciones monotónicas, donde las variables pueden cambiar conjuntamente sin necesidad de ser lineales. Es robusto frente a valores atípicos y no requiere que los datos o sus residuos sigan una distribución normal, lo que lo hace más apropiado para situaciones en las que las etiquetas son ordinales o cuando la relación entre las variables no es estrictamente lineal.

Los resultados muestran que la distribución de las correlaciones respecto a cada característica es similar en ambos métodos. Sin embargo, la prosodia presenta la mayor correlación con las etiquetas de *EpaDB*, lo que sugiere que el modelo captura mejor las variaciones prosódicas en relación con la natividad.

Visualización mediante Gráficos de Dispersión Para complementar el análisis numérico de correlación y visualizar de manera más intuitiva la relación entre las variables, se generaron gráficos de dispersión para cada característica en relación con el puntaje de natividad.

Como ejemplo, la Figura 3.9 muestra el gráfico de dispersión correspondiente al hablante 17, ilustrando la relación entre el puntaje de prosodia por frase y el grado de natividad.

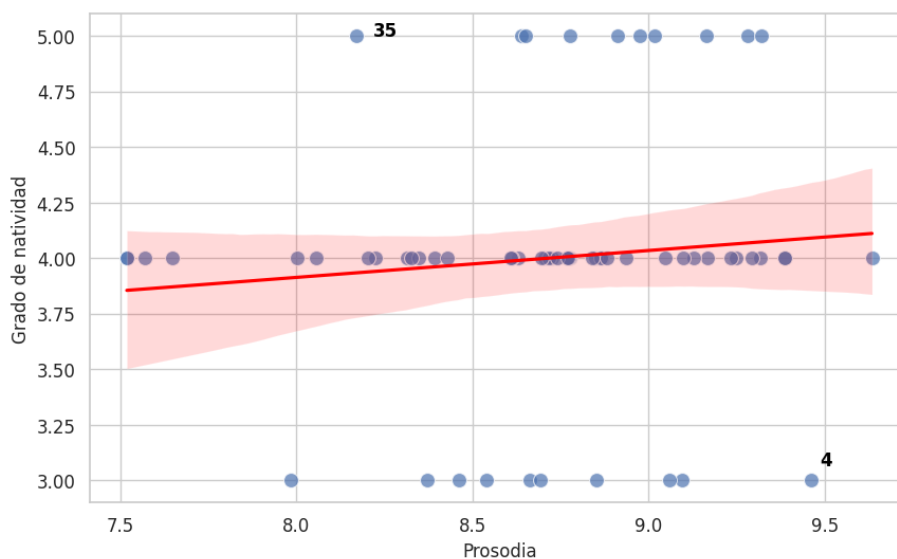


Figura 3.9: Gráfico de dispersión para el hablante 17, grado de natividad en función de la prosodia.

En este gráfico, cada punto representa un audio individual, en algunos casos identificado por un número que facilita el seguimiento y análisis posterior. Se incluye además una línea de tendencia obtenida mediante regresión lineal, con el fin de visualizar la dirección y fuerza de la relación entre las variables.

Observando la Figura 3.9, se aprecia una ligera tendencia ascendente, consistente con el coeficiente de correlación positivo, pero débil, obtenido anteriormente. No obstante, también se identifican varios puntos que se desvían significativamente de la línea de tendencia, indicando la presencia de valores atípicos.

Los valores atípicos identificados en los gráficos de dispersión representan audios cuyas puntuaciones no siguen el patrón general observado entre cada característica y el puntaje de natividad. Para comprender las posibles causas de estas desviaciones, se procedió a un análisis cualitativo detallado de estos casos específicos.

Dicho análisis implicó la escucha manual de los audios correspondientes a los valores atípicos y se presenta en la sección 3.3.2.

El análisis de correlación, respaldado por la visualización mediante gráficos de dispersión, indica que las características extraídas por el modelo no presentan una asociación fuerte con el puntaje de natividad. Esto sugiere que el modelo puede estar omitiendo factores importantes o que las características actuales no capturan adecuadamente la complejidad de la natividad en el habla.

Análisis de Componentes Principales (PCA)

Se realizó con el objetivo de explorar la capacidad de las características evaluadas por el modelo para diferenciar entre los distintos grados de natividad de los hablantes. Este análisis sirve como una confirmación adicional de los resultados obtenidos en los análisis descriptivo básico y de correlación, que ya sugerían limitaciones en estas características para discriminar eficazmente entre los grados de natividad.

Para determinar el número óptimo de componentes principales a retener, se analizó la varianza explicada acumulada por cada componente, como se muestra en la figura 3.10. El gráfico revela que el primer componente principal explica aproximadamente el 96.5 % de la variabilidad total en los datos. Al agregar el segundo componente, se incrementa la varianza explicada acumulada a aproximadamente el 99 %. Aunque los componentes adicionales continúan aumentando la varianza explicada, su contribución marginal es pequeña, alcanzando casi el 100 % con los cuatro componentes.

Dado que los dos primeros componentes principales explican en conjunto más del 99 % de la variabilidad total, se decidió retener estos dos componentes para el análisis posterior. Este criterio permite capturar la mayor parte de la información relevante contenida en las características evaluadas, mientras se simplifica el modelo reduciendo la dimensionalidad.

Una vez seleccionados los componentes principales, se proyectaron los datos en el espacio definido por los dos primeros. La figura 3.11 muestra la distribución de los datos en este espacio bidimensional, con los puntos coloreados según

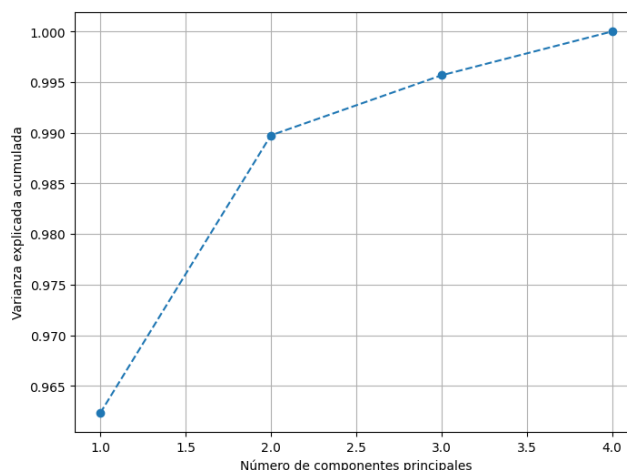


Figura 3.10: Varianza explicada acumulada por cada componente principal.

el grado de natividad. Este gráfico proporciona una visualización de cómo se distribuyen los datos en función de los componentes principales y si es posible observar alguna tendencia de agrupamiento según el grado de natividad.

El resultado de la proyección en los componentes principales confirma lo que ya se había anticipado en los análisis descriptivo y de correlación: no se observa una separación clara entre los diferentes grados de natividad en el espacio bidimensional formado por los dos primeros componentes principales. Los puntos correspondientes a cada puntaje se encuentran distribuidos de manera superpuesta, lo que refuerza la idea de que las características evaluadas no son suficientes para diferenciar de manera efectiva entre los grados de natividad.

Esta superposición, junto con los resultados de los análisis previos, sugiere que estas características, en su forma actual, no proporcionan la información necesaria para predecir el grado de natividad de manera precisa.

Análisis cualitativo

El análisis cualitativo se centró en comprender las discrepancias observadas entre las evaluaciones automáticas del modelo *MultiPA* y las etiquetas de natividad asignadas por *EpaDB*. A través de una escucha detallada, se investigaron posibles razones detrás de estas discrepancias, poniendo especial atención en cómo ciertas características del habla pudieron haber influido en la evaluación automática.

Para identificar los casos más relevantes, se seleccionaron los valores atípicos a partir de los gráficos de dispersión generados en la sección de análisis de correlación. La **Tabla 3.5** resume algunos outliers, mostrando las frases correspondientes, junto con las etiquetas de *EpaDB* y los puntajes otorgados por *MultiPA*. Aunque se analizaron más casos, esta sección presenta ejemplos

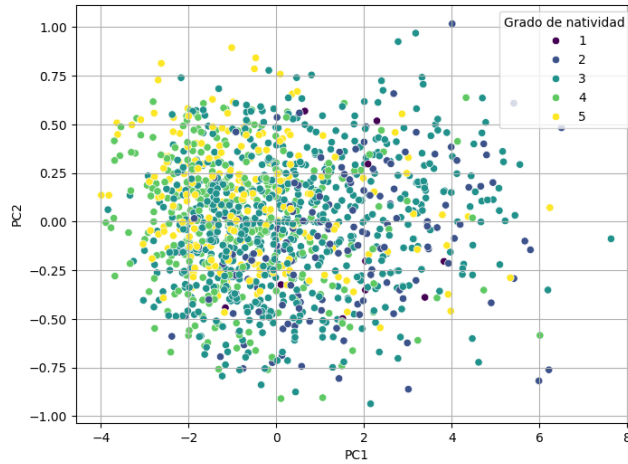


Figura 3.11: Proyección en los dos primeros componentes principales según el grado de natividad.

representativos que destacan las discrepancias más significativas entre las evaluaciones automáticas y humanas.

ID	Frase	EpaDB	Precisión	Fluidez	Prosodia	Total
14.19	"My best friend speaks English and Spanish"	1	8.71	9.23	8.70	8.55
14.13	"I had a headache last year"	3	6.99	7.98	7.19	6.91
14.45	"Put your coat on"	3	7.03	7.51	7.05	6.90
15.50	"My family will visit Argentina in July"	2	8.80	9.25	8.78	8.59
15.73	"I'm sure it's okay"	2	8.76	9.53	9.14	8.88
17.04	"My father is at home."	3	9.27	9.82	9.46	9.48
17.35	"They said so on the news."	5	8.18	8.76	8.17	8.31

Tabla 3.5: Comparación de outliers entre *MultiPA* y *EpaDB*.

En la columna "ID", el primer número corresponde al identificador del hablante, mientras que el segundo número se refiere al identificador del archivo de audio.

En el **audio 19 del hablante 14**, correspondiente a la frase "My best friend speaks English and Spanish", *EpaDB* asignó un grado de 1, mientras que *MultiPA* otorgó un puntaje total de 8.55. Este caso destaca como un *outlier* porque, a pesar de la baja calificación de natividad por parte de *EpaDB*, *MultiPA* evaluó el audio con puntajes altos en todas las características. El acento español rioplatense es evidente, lo que probablemente influyó en la calificación de *EpaDB*, mientras que *MultiPA* priorizó la inteligibilidad y la *fluidez* del discurso. Al escuchar el audio, la frase se entiende claramente y de forma fluida, lo que justifica la alta puntuación de *MultiPA*. Sin embargo, se infiere que el acento claramente no nativo fue lo que llevó a *EpaDB* a otorgar el valor mínimo.

En contraste, el **audio 13 del mismo hablante**, con la frase "I had a headache last year", recibió una etiqueta de 3 en *EpaDB*, mientras que *MultiPA*

otorgó puntajes significativamente más bajos en comparación con el audio 19. Aunque en este caso el acento no es tan evidente, la pronunciación es claramente más deficiente, lo que afecta la inteligibilidad del discurso. Al escuchar el audio, se perciben errores en la pronunciación que comprometen la calidad general del habla, lo que explica las calificaciones más bajas de *MultiPA*. Sin embargo, *EpaDB* otorgó una calificación más alta, lo que demuestra una *inversión* en las evaluaciones: la puntuación de *EpaDB* fue menos estricta en la ausencia de un acento notorio, mientras que *MultiPA* penalizó más la falta de precisión en la pronunciación.

En el **audio 45 del hablante 14**, donde se repite la palabra “put” en la frase “Put your coat on”, *MultiPA* asignó un puntaje de 6.9, reflejando la disfluencia observada, mientras que *EpaDB* le otorgó un grado de natividad de 3. Este caso sugiere que *MultiPA* podría percibir las interrupciones en la fluidez de manera más rigurosa que los evaluadores humanos de esta base.

En cuanto al **hablante 15**, se compararon dos audios. En el **audio 50**, correspondiente a la frase “My family will visit Argentina in July”, la palabra “Argentina” fue pronunciada en español. *EpaDB* asignó una etiqueta de 2, probablemente influenciado por esta pronunciación, mientras que *MultiPA* evaluó el audio con un puntaje total de 8.59, destacando la claridad general del discurso. Sin embargo, el **audio 73**, correspondiente a la frase “I’m sure it’s okay”, también recibió un 2 en *EpaDB*, pero desde una evaluación auditiva personal, este último audio parece mejor pronunciado y suena más cercano a un hablante nativo en comparación con el audio 50. *MultiPA* otorgó al audio 73 un puntaje total de 8.88, con altos valores en fluidez y prosodia. A pesar de que ambos audios recibieron la misma etiqueta de *EpaDB*, *MultiPA* presenta una distinción entre ellos, lo que sugiere que el modelo prioriza la claridad técnica y la fluidez, mientras que *EpaDB* podría estar considerando otros factores más sutiles relacionados con la natividad.

Finalmente analizamos los **audios 4 y 35 del hablante 17**. En la figura 3.9 se pueden ver identificados como outliers respecto al valor de prosodia.

En el caso del audio 4, con la frase “My father is at home”, los valores de *MultiPA* son significativamente altos, mientras que el valor de natividad es medio (3). Al escuchar el audio, la pronunciación se considera muy buena a excepción de la palabra “father”, cuya “th”, representada por el fonema /ð/ en el Alfabeto Fonético Internacional (IPA), fue sustituida por una /d/. Esta sustitución de fonemas es una característica común en hablantes nativos de español, dado que el fonema /ð/ no existe en el conjunto de fonemas del español y, en su lugar, se reemplaza por el fonema más cercano disponible, que es /d/. Esta sustitución podría ser la razón por la cual el grado de natividad en *EpaDB* es medio. Respecto a los puntajes de *MultiPA*, el menor es el de precisión, lo cual concuerda con la apreciación realizada.

El audio 35 recibió un puntaje en *EpaDB* de 5, mientras que los puntajes en *MultiPA* son medios. De todas formas, al escuchar el audio se considera que la pronunciación es correcta, sin errores evidentes en las palabras o patrones prosódicos.

Conclusiones del Análisis Cuantitativo y Cualitativo El análisis tanto cuantitativo como cualitativo revela que *MultiPA*, aunque evalúa aspectos como la claridad, fluidez y prosodia, no distingue grados de natividad tal como lo consideran en *EpaDB*. En el análisis cuantitativo, los resultados indicaron que el modelo no puede diferenciar claramente entre los distintos niveles de natividad asignados. El análisis cualitativo complementa esta observación al mostrar que *MultiPA* asigna puntajes altos incluso en audios con acentos marcados, sugiriendo que el modelo no capta algunos de los factores que influyen en la evaluación humana de la natividad.

Es importante destacar que la base de datos *EpaDB* fue utilizada porque proporciona una etiqueta a nivel de frase que podría estar vinculada con aspectos de pronunciación. Aunque el objetivo no es evaluar directamente la natividad, esta etiqueta permitió explorar si el modelo *MultiPA* puede detectar diferencias en pronunciación basadas en ella.

En ambos análisis, se sugiere que la natividad involucra sutilezas fonéticas complejas, como la precisión fonética y el acento, que *MultiPA* no parece captar de manera efectiva. Esto resalta la dificultad de automatizar la evaluación de características del habla cuando los elementos más finos, como el acento, son determinantes en la percepción humana. Sin embargo, esto no implica que *MultiPA* no sea útil; simplemente significa que las características que evalúa no están directamente relacionadas con la natividad. Esto puede tener sentido si consideramos que los autores de *MultiPA* probablemente están midiendo la pronunciación desde una perspectiva más centrada en la inteligibilidad, en lugar de la natividad.

A continuación, se presenta un análisis utilizando la base de datos de Ceibal, que es particularmente relevante al ser el caso de uso de este estudio. Dado que la evaluación de Ceibal se centra en la inteligibilidad, también resulta interesante el análisis para explorar si *MultiPA* está más alineado con esta definición de pronunciación, ya que vimos anteriormente que no parece estar significativamente relacionado con la natividad.

Evaluación sobre base de Ceibal

En el caso de la base de datos de Ceibal, descrita en la Sección 3.2.1, se utilizará el puntaje holístico de la categoría *Delivery*, ya que en esta métrica los evaluadores consideran aspectos como inteligibilidad y fluidez. El puntaje de *Delivery* se asigna en una escala de 0 a 3, donde 0 representa una baja inteligibilidad y fluidez, y 3 indica una alta calidad en ambos aspectos.

A continuación, se presenta un histograma con la distribución de los puntajes asignados por los evaluadores humanos (Figura 3.12). El histograma presentado muestra la distribución de los puntajes de *Delivery* asignados por los evaluadores humanos en esta base de datos. Se observa que la mayoría de los estudiantes tienen un puntaje de 1, lo que indica que estos fueron evaluados con inteligibilidad y fluidez moderada. Solo un número reducido recibió puntajes de 2 o 3, sugiriendo que la mayoría de los estudiantes presentaron dificultades en la fluidez e inteligibilidad de su expresión oral.

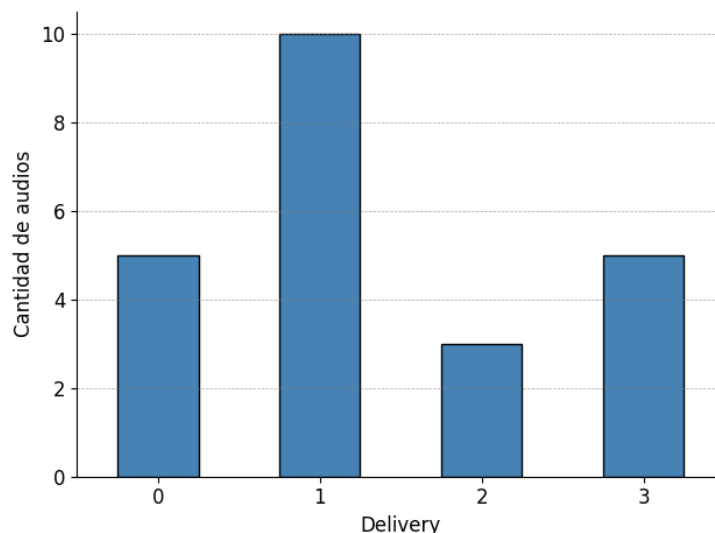


Figura 3.12: Distribución de los puntajes de Delivery en la base de datos de Ceibal.

La inferencia del modelo *MultiPA* sobre los audios de la base de datos de Ceibal no fue un proceso directo, ya que fue necesario realizar un preprocesamiento para garantizar la calidad de los datos de entrada. En primer lugar, se utilizó un sistema de reconocimiento automático del habla para transcribir los audios, seguido de un detector de idiomas que permitió identificar aquellos que contenían habla en inglés. Los audios que no cumplían con este criterio fueron descartados. Luego, los audios seleccionados fueron segmentados para eliminar silencios prolongados al inicio y al final, lo cual redujo significativamente su duración en muchos casos.

Debido a este preprocesamiento, no todos los hablantes mantienen los 14 audios originalmente previstos. Además, algunos audios presentaron errores durante la ejecución de *MultiPA*, lo que también contribuyó a la reducción del número de audios procesados por hablante. Como resultado, la cantidad de audios finales varía entre estudiantes.

Dado que las etiquetas de *Delivery* reflejan una evaluación global por hablante, no es posible realizar una comparación directa con los resultados de *MultiPA*, que evalúa cada audio individualmente. Por este motivo, se calcularon métricas agregadas, como la media y la mediana de las cuatro características evaluadas por *MultiPA* para cada hablante.

Análisis cuantitativo

El objetivo de esta sección es explorar cómo las características evaluadas por *MultiPA* se relacionan con las etiquetas globales de Delivery. Para ello, se

realizará la comparación entre las métricas agregadas calculadas a partir de los resultados de la inferencia en *MultiPA* de los audios, y las etiquetas globales.

Análisis descriptivo

En esta sección, se realiza un análisis descriptivo de las características evaluadas por *MultiPA* para los hablantes de la base de datos de Ceibal, utilizando métricas agregadas como la media y la mediana. El objetivo de este análisis es observar cómo varían las puntuaciones asignadas a cada hablante y evaluar si estas métricas agregadas ofrecen una representación coherente con las evaluaciones globales de *Delivery*. A su vez, se busca detectar posibles valores atípicos o anomalías en las puntuaciones. En la sección A.2 del anexo se incluyen las tablas asociadas al análisis descriptivo.

Como parte del análisis, se generaron dos tipos de diagramas de caja para cada característica: uno con las puntuaciones agregadas utilizando la media y otro utilizando la mediana. En ambos casos, las etiquetas globales de *Delivery* en el eje x representan la clasificación asignada a cada hablante por los evaluadores humanos. Estos diagramas permiten visualizar la centralidad, dispersión y valores extremos en las puntuaciones asignadas por *MultiPA*. Dado que ambos enfoques muestran tendencias similares, se optó por presentar únicamente los diagramas basados en la media (ver Figura 3.13).

En los diagramas se observa una clara tendencia general de incremento en las puntuaciones de las características conforme aumenta la etiqueta global de *Delivery*. A medida que el *Delivery* pasa de 0 a 2, las medianas de todas las características aumentan progresivamente, lo que indica que, en general, *MultiPA* asigna mayores puntuaciones a hablantes con *Delivery* más alto.

Sin embargo, en *Delivery 3* se observa una leve disminución en la mediana, que podría deberse a imperfecciones en los datos o a variabilidad en la evaluación de los audios. Cabe recordar que estos resultados provienen de métricas agregadas (media) y que la cantidad de audios por hablante varía debido al preprocesamiento y a los errores ocurridos durante la ejecución de *MultiPA*.

A pesar de la disminución en la media, destaca un *outlier* en *Delivery 3* que sigue la tendencia esperada de puntuaciones más altas con mayor *Delivery*, lo que sugiere que algunos hablantes aún alcanzan buenos resultados, aunque la tendencia general se vea afectada.

Análisis de correlación

Se calcularon los coeficientes de correlación de Pearson y Spearman entre las etiquetas globales de *Delivery* de la base de datos de Ceibal y las métricas agregadas de precisión, fluidez, prosodia y puntaje total evaluadas por *MultiPA*, tanto para la media como para la mediana.

En la Tabla 3.6 se presentan los coeficientes de Pearson calculados con la base de datos de Ceibal, mientras que en la Tabla 3.7 se muestran los coeficientes de Spearman.

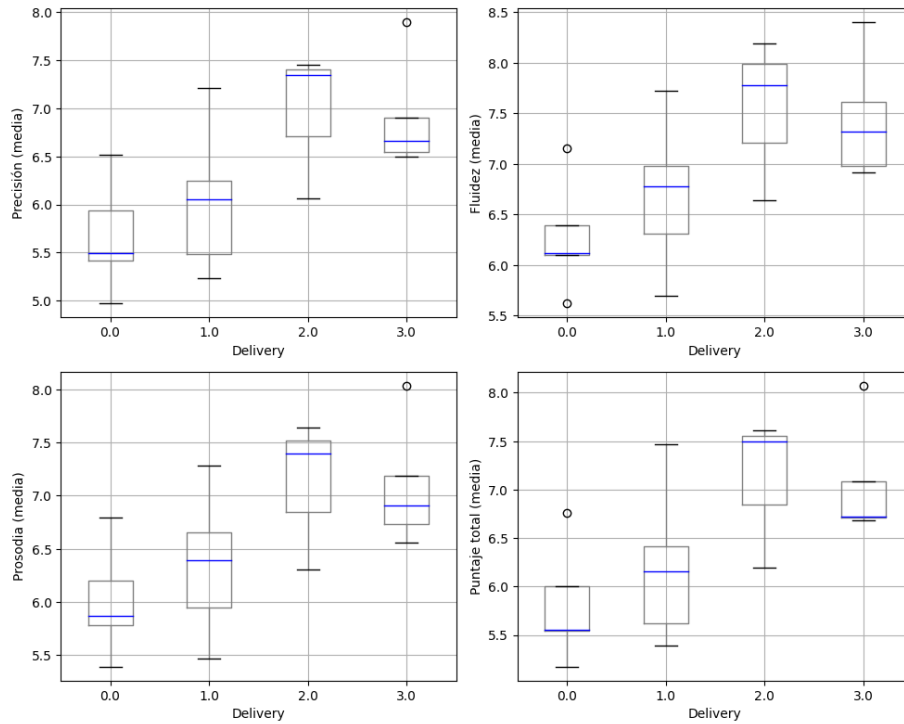


Figura 3.13: Diagramas de caja de las características evaluadas por *MultiPA* utilizando la media como métrica agregada.

Métrica	Precisión	Fluidez	Prosodia	Puntaje Total
Pearson	0.635	0.611	0.618	0.629
-Media-	(1.13×10^{-3})	(1.96×10^{-3})	(1.66×10^{-3})	(1.31×10^{-3})
Pearson	0.551	0.509	0.513	0.574
-Mediana-	(6.47×10^{-3})	(1.31×10^{-2})	(1.23×10^{-2})	(4.21×10^{-3})

Tabla 3.6: Coeficientes de Pearson y valores p (en notación científica) para las métricas agregadas (media y mediana) en la base de datos de Ceibal.

Métrica	Precisión	Fluidez	Prosodia	Puntaje Total
Spearman	0.654	0.608	0.627	0.632
-Media-	(7.17×10^{-4})	(2.09×10^{-3})	(1.38×10^{-3})	(1.22×10^{-3})
Spearman	0.553	0.559	0.547	0.594
-Mediana-	(6.26×10^{-3})	(5.58×10^{-3})	(6.94×10^{-3})	(2.82×10^{-3})

Tabla 3.7: Coeficientes de Spearman y valores p (en notación científica) para las métricas agregadas (media y mediana) en la base de datos de Ceibal.

Los coeficientes de correlación muestran una relación positiva entre las métricas agregadas y las etiquetas globales de *Delivery*. Las correlaciones calculadas para la media de las métricas son ligeramente superiores a las obtenidas para la mediana, lo que sugiere que la media captura mejor la relación entre los resultados de *MultiPA* y las etiquetas globales.

La métrica de precisión destaca como la que presenta las correlaciones más altas respecto a la media, tanto en Pearson como en Spearman, lo que indica que esta característica podría ser particularmente relevante para la evaluación de inteligibilidad y fluidez en las respuestas. Para el caso de la mediana, el puntaje total es el que tiene mayor correlación, aunque siendo levemente menor al obtenido utilizando la media como métrica agregada. Los valores p asociados, menores a 0.01 en todos los casos, confirman que las correlaciones observadas son estadísticamente significativas.

Dado que la evaluación en la base de datos se centra en aspectos holísticos como la inteligibilidad y el *flow of speech*, tiene sentido complementar el análisis de las métricas individuales con una **métrica agregada global**. Esta métrica combina las medias (o medianas) de las características evaluadas por MultiPA, uniendo precisión, fluidez, prosodia y puntaje total en un solo valor representativo. El objetivo de esta agregación no es necesariamente mejorar las correlaciones con las etiquetas de *Delivery*, sino aproximar el enfoque integrado de los evaluadores humanos, quienes probablemente consideran múltiples aspectos de la pronunciación al emitir su juicio.

En la Tabla 3.8 se presentan los coeficientes de correlación de Pearson y Spearman, junto con sus valores p, para esta métrica agregada global.

Métrica agregada global	Pearson	Spearman
Media de las medias	0.627 (1.36×10^{-3})	0.625 (1.43×10^{-3})
Mediana de las medianas	0.558 (5.70×10^{-3})	0.582 (3.56×10^{-3})

Tabla 3.8: Coeficientes de Pearson y Spearman con valores p (en notación científica) para la métrica agregada global (media de las medias y mediana de las medianas).

Aunque las correlaciones obtenidas son similares a las de las métricas individuales, la agregación de las características proporciona una representación más general. No mejora significativamente las correlaciones respecto a las métricas por separado, pero tampoco afecta negativamente la relación con las etiquetas de *Delivery*. Por lo tanto, es una opción válida para representar las características globales, alineándose con el enfoque holístico utilizado por los evaluadores humanos.

Correlación entre la cantidad de audios y el puntaje de *Delivery*

En el pre-procesamiento de esta base de datos, se retuvieron únicamente

los audios que contenían algo de contenido en inglés, descartando aquellos que no lo tenían por no ser relevantes para el análisis. Como resultado, el número de audios utilizados por estudiante para la inferencia en *MultiPA* varía. Esto implica que los estudiantes con menos audios asociados son aquellos que no lograron producir suficientes respuestas en inglés, lo que podría influir en su puntaje de *Delivery*.

Para evaluar si esta variabilidad en la cantidad de audios afecta el puntaje, se calculó la correlación de *Spearman* entre el número de audios en inglés y el puntaje de *Delivery*. Los resultados, presentados en la Tabla 3.9, muestran una correlación de **0.207**, lo que indica una relación positiva pero muy débil entre ambas variables.

El valor p asociado fue de **0.342**, lo que significa que no hay suficiente evidencia estadística para afirmar que esta relación es significativa. En otras palabras, aunque parece haber una ligera tendencia a que más audios en inglés estén relacionados con un mayor puntaje de *Delivery*, este resultado no es lo suficientemente fuerte como para sacar conclusiones definitivas.

	Correlación de Spearman	valor p
Cantidad de audios vs. <i>Delivery</i>	0.207	0.342

Tabla 3.9: Correlación de Spearman entre la cantidad de audios y *Delivery*

Visualización mediante gráfico de dispersión

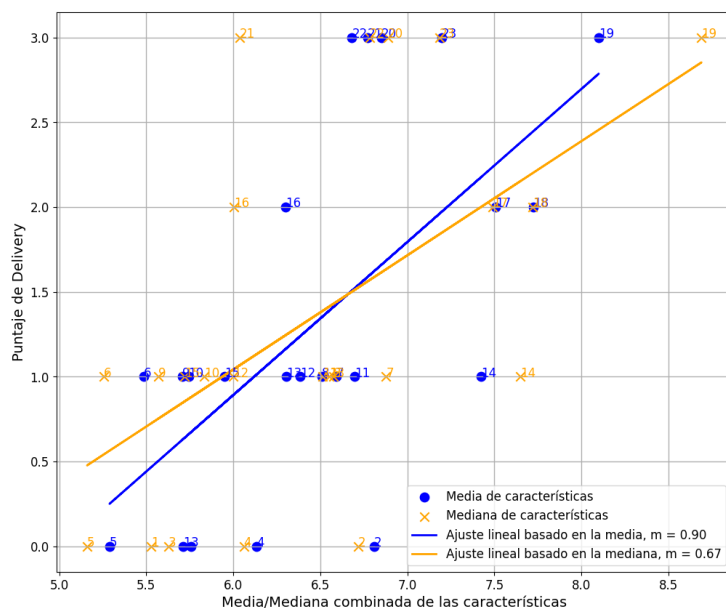
El gráfico de dispersión presentado a continuación (Figura 3.14) muestra la relación entre las métricas agregadas (media y mediana combinadas de las características evaluadas por *MultiPA*) y el puntaje de *Delivery* asignado a cada estudiante. El propósito del gráfico es complementar el análisis cuantitativo previo, proporcionando una representación visual que ilustre la relación entre las métricas agregadas y las etiquetas de *Delivery*, y mostrando cómo se distribuyen los hablantes. En el *eje X* se representan estas métricas agregadas, mientras que en el *eje Y* se muestra el puntaje de *Delivery*.

Cada punto en el gráfico corresponde a un estudiante, y su posición refleja la media o la mediana combinada de las características. Los puntos están coloreados para diferenciar entre las medias (en azul) y las medianas (en naranja). Además, sobre cada punto se coloca un número, que corresponde a un identificador de estudiante generado internamente, con el objetivo de facilitar la identificación en un análisis posterior.

En el gráfico se trazan dos rectas de ajuste lineal: una para la media y otra para la mediana combinada. La pendiente de la recta de ajuste para la media es de 0.90, mientras que la pendiente de la mediana es de 0.67. Esto refuerza lo observado en el análisis previo, donde la métrica agregada basada en la media mostró una correlación ligeramente mayor con el puntaje de *Delivery* en comparación con la mediana. Ambas pendientes son positivas, lo que indica que, a medida que aumentan las características agregadas, también lo hace el

puntaje de *Delivery*.

Las rectas de ajuste lineal refuerzan la coherencia de estas tendencias, demostrando que las métricas agregadas pueden representar de manera global el rendimiento en pronunciación, alineándose con el enfoque holístico utilizado por los evaluadores humanos.



Caso estudiante Id. 2

En el caso del estudiante identificado con el número 2, el puntaje de *Delivery* es 0 mientras que la media y mediana de las características inferidas por MultiPA, son 6.8 y 6.71 respectivamente. Esto lo hace un *outlier* ya que en otros estudiantes con puntaje de *Delivery* 0, la media y mediana tienden a oscilar entre 5 y 6.5.

Esta discrepancia puede deberse a varios factores observados en los audios procesados. En primer lugar, solo 5 de los 14 audios del estudiante fueron clasificados como audios en inglés y, por lo tanto, procesados mediante MultiPA, mientras que los 9 restantes fueron excluidos del análisis. Al escuchar estos 5 audios se nota que la pronunciación es clara y el habla fluída, aunque en algunos casos el contenido de las respuestas es confuso y no siempre coherente, por ejemplo debido a estructuras gramaticales incorrectas. Estas características de la forma de producción del habla pueden haber contribuido a los valores altos otorgados por MultiPA en sus métricas.

Otro aspecto que influye en la interpretación es la repetición de la consigna de la pregunta en algunas respuestas, un factor que los evaluadores humanos suelen penalizar, ya que refleja una falta de respuesta original. Sin embargo, nosotros no tuvimos en cuenta este factor en la evaluación automática, lo cual puede ser otra de las razones de la discrepancia entre el puntaje de *Delivery* humano y las métricas obtenidas mediante el modelo.

Caso estudiante Id. 14

Para el estudiante con identificador 14, el puntaje de *Delivery* asignado es 1, mientras que la media y mediana de las características inferidas por MultiPA son más altas, alcanzando valores de 7.42 y 7.65 respectivamente. Este caso se considera un *outlier* porque muestra un puntaje superior en comparación con otros estudiantes cuya etiqueta de *Delivery* es 1.

En este caso, 12 de los 14 audios fueron clasificados como en inglés y procesados por MultiPA, mientras que los 2 restantes fueron excluidos. Entre los audios en inglés, hay 3 en particular que contribuyen significativamente a los puntajes altos observados. El primero de estos audios incluye la lectura de la consigna, lo cual, como se mencionó anteriormente, puede hacer que la evaluación humana lo considere como una limitación pero el modelo automático no es capaz de diferenciar. Sin embargo, los otros dos audios muestran un nivel elevado de *Delivery*, con pronunciación clara y buena fluidez, características que probablemente han influido en el puntaje elevado otorgado por MultiPA en este caso.

Caso estudiante Id. 21

Para el estudiante con identificador 21, el puntaje de *Delivery* asignado es 3 (el máximo), mientras que la media y mediana de las características inferidas por MultiPA son 6.77 y 6.03 respectivamente. En este caso, es la mediana la medida que lo clasifica como un *outlier*, dado su valor inferior en comparación con otros estudiantes con puntaje de *Delivery* 3.

Este estudiante tiene 12 audios clasificados como en inglés y 2 que no. En particular, hay 2 audios que presentan puntajes significativamente bajos. En

ambos casos, el nivel de ruido en las respuestas es elevado, lo que requiere un esfuerzo humano considerable para comprender lo que el estudiante está diciendo. Sin embargo, al final, el contenido y la forma de producción del habla en las respuestas es correcto, lo cual explica el buen puntaje asignado en *Delivery*. Se infiere entonces que la razón de los puntajes bajos en estos audios es el alto nivel de ruido, que afecta la evaluación automática realizada por MultiPA.

3.4. Conclusiones del Capítulo

En este capítulo se exploraron diversos métodos para la evaluación automática de la pronunciación en el aprendizaje de segundas lenguas. Se analizó cómo la percepción de lo que constituye una buena pronunciación ha evolucionado, pasando de un énfasis en sonar como un hablante nativo a priorizar la inteligibilidad y la comprensibilidad.

Se implementaron dos enfoques principales: una línea base basada en la comparación fonética y el modelo avanzado *MultiPA*. La línea base mostró limitaciones significativas al correlacionar con los puntajes de natividad en la base *EpaDB*, lo que evidencia la dificultad de este método para captar las sutilezas de la pronunciación que influyen en la percepción humana.

Por otro lado, *MultiPA* fue evaluado tanto en *EpaDB* como en datos de Ceibal. En *EpaDB*, *MultiPA* no logró correlaciones fuertes con los puntajes de natividad, probablemente porque este modelo está orientado a evaluar la inteligibilidad más que la similitud con hablantes nativos. Sin embargo, en la base de Ceibal, los resultados fueron particularmente prometedores. *MultiPA* mostró buenas correlaciones con las evaluaciones humanas en la categoría de *Delivery*, que se enfoca en la inteligibilidad y la fluidez del discurso.

Estos hallazgos sugieren que *MultiPA* es adecuado para contextos donde el objetivo principal es mejorar la comunicación efectiva, alineándose con enfoques modernos de enseñanza que priorizan la comprensibilidad sobre la perfección fonética nativa. Además, la capacidad de *MultiPA* para trabajar con respuestas abiertas y su desempeño positivo en datos reales de estudiantes hispanohablantes refuerzan su potencial utilidad en programas como *Ceibal en Inglés*.

Capítulo 4

Evaluación de Características Semánticas

4.1. Introducción a la Evaluación Semántica

Como ya se ha mencionado, la evaluación del habla en un segundo idioma es un proceso multifacético que abarca diversos aspectos, desde la forma de producción del habla hasta la coherencia y la adecuación del contenido. En este capítulo, se aborda un componente importante y a menudo complejo: la evaluación semántica de las respuestas orales. La evaluación semántica se refiere al análisis del significado y la pertinencia del contenido expresado por el hablante, más allá de la forma lingüística utilizada [48].

El proceso de evaluar el contenido semántico de las respuestas orales presenta desafíos particulares en el contexto del aprendizaje de un segundo idioma. La tarea de analizar no solo cómo se expresa el estudiante, sino también qué comunica, requiere un enfoque capaz de interpretar y analizar el significado de lo transmitido en el audio. Este proceso es especialmente complejo dado que las respuestas orales a menudo contienen matices, ambigüedades y estructuras lingüísticas que pueden ser difíciles de evaluar de manera automatizada.

Los enfoques tradicionales para la evaluación semántica han dependido en gran medida de evaluadores humanos, lo que puede resultar costoso y poco escalable. Sin embargo, los avances recientes en el procesamiento del lenguaje natural y, en particular, en los modelos de lenguaje de gran escala (LLMs, por sus siglas en inglés) han abierto nuevas posibilidades para la automatización de este proceso.

En este capítulo, se explora un enfoque para la evaluación semántica automatizada, centrándose específicamente en la evaluación de la correctitud de las tareas en el marco de las pruebas de habla en inglés.

En las secciones siguientes, se exploran detalladamente los componentes de este enfoque:

- La definición y los desafíos del *Task Fulfillment* en el contexto de la eva-

luación del habla en inglés.

- La base de datos utilizada para el análisis.
- La arquitectura propuesta para el sistema de evaluación automática.
- La selección y evaluación de sistemas de Reconocimiento Automático del Habla (ASR).
- El diseño e implementación del analizador semántico, basado en Modelos de Lenguaje de Gran Escala.
- La metodología de evaluación del sistema y los resultados obtenidos.

4.2. *Task Fulfillment*

4.2.1. Definición e Importancia

El concepto de *Task Fulfillment* o *Cumplimiento de la Tarea* es fundamental en la evaluación del habla en un segundo idioma. Se refiere a la capacidad del estudiante para responder adecuadamente a una pregunta o completar una tarea oral de manera que aborde todos los aspectos requeridos. Este componente de la evaluación va más allá de la corrección gramatical o de la forma de producción del habla, centrándose en la relevancia y completitud del contenido proporcionado por el estudiante.

La importancia del task fulfillment radica en su capacidad para evaluar la competencia comunicativa real del estudiante. No solo se trata de hablar correctamente, sino de transmitir efectivamente la información solicitada o requerida en un contexto específico. Esta habilidad es muy relevante en situaciones de la vida real donde el uso del idioma va más allá de la mera producción de frases gramaticalmente correctas.

4.2.2. Categorización de Tareas

La prueba de *Ceibal en Inglés* se compone de 4 partes tal y como se detalla en la sección 2.3. Destacar que las Partes 1 y 3 comprenden preguntas vinculadas a la experiencia personal del estudiante, mientras que las Partes 2 y 4 se centran en evaluar la capacidad de observar, interpretar y describir en inglés un contexto visual específico. Por lo tanto, teniendo en cuenta esta observación, hemos categorizado las tareas de estas partes en dos grupos principales como se detalla a continuación:

Pregunta / Respuesta sin contexto de referencia. Esta involucra la parte uno y tres de la prueba cuyas preguntas son del estilo “What’s your name?” o “Tell me about your family”. En este caso, la evaluación se centra en la coherencia y relevancia de la respuesta a la pregunta planteada.

Pregunta / Respuesta con contexto de referencia. Esta involucra la parte 2 y 4 de la prueba. Aquí, la evaluación debe considerar no solo la respuesta en sí, sino también su ajuste al contexto proporcionado.

Esta categorización permite simplificar el problema y desarrollar estrategias específicas para cada tipo de tarea. Sin embargo, la evaluación automatizada del *task fulfillment* presenta varios desafíos específicos que subrayan la complejidad de este proceso.

La **variabilidad en las respuestas** a preguntas abiertas representa uno de los principales desafíos, dificultando la definición de criterios de evaluación estrictos. Las respuestas de los estudiantes pueden presentar diversidad no solo en contenido y estructura, sino también en aspectos lingüísticos propios de hablantes no nativos, comúnmente llamados *English as a Second Language (ESL) variations*, incluyendo variaciones gramaticales, léxicas y sintácticas. Esto requiere un **sistema de evaluación flexible**, capaz de comprender el significado pretendido más allá de los errores de expresión, y evaluar de manera comprensiva una amplia gama de manifestaciones lingüísticas.

En el caso de **tareas basadas en imágenes**, el desafío se amplía. El sistema no solo debe comprender la respuesta verbal del estudiante, sino también relacionarla con el contexto visual proporcionado. Esto implica la necesidad de desarrollar mecanismos que puedan “interpretar” la relación entre el lenguaje y la información visual.

Otro reto significativo es la **evaluación de la completitud** de las respuestas, especialmente en tareas que requieren elaboraciones más extensas. Determinar si una respuesta aborda todos los aspectos requeridos de una tarea implica un análisis detallado del contenido, considerando tanto la información explícita como la implícita en la respuesta del estudiante.

Estos desafíos resaltan la necesidad de un enfoque que pueda capturar los matices de las respuestas de los estudiantes. El trabajo realizado se centra en abordar estos retos, desarrollando un sistema capaz de evaluar el *task fulfillment* en una variedad de respuestas y tipos de tareas.

4.3. Base de Datos para la Evaluación de Task Fulfillment

Para desarrollar y evaluar el sistema de evaluación automatizada del habla, se construyó una base de datos propia inspirada en los tipos de ejercicios comúnmente utilizados en evaluaciones orales de inglés, como la prueba de *Ceibal en Inglés* (2.3). Esta base de datos intenta, a pequeña escala, reflejar la diversidad de tareas y respuestas que se encuentran en evaluaciones reales.

La base de datos incluye un total de **35 tareas** diferentes, abarcando los cuatro tipos principales identificados en la prueba de *Ceibal en Inglés*:

- **Tipo 1 - Pregunta Simple con Respuesta Libre (16 tareas):** Consiste en preguntas directas que requieren una respuesta personal y abierta.

Por ejemplo, “Can you swim?” o “What is your name?”. Estas preguntas evalúan la capacidad del estudiante para comprender y responder a consultas básicas en inglés.

- **Tipo 2 - Pregunta Contextual con Respuesta Específica (9 tareas):** Abarca preguntas basadas en un contexto dado, generalmente visual. Ejemplos incluyen “What food is it?”. Estas tareas evalúan la habilidad del estudiante para interpretar información visual y describirla en inglés.
- **Tipo 3 - Tarea Guiada con Respuesta Elaborada (9 tareas):** Corresponde a tareas más amplias que proporcionan puntos de guía para estructurar una respuesta más extensa, como “Tell me about your holidays” acompañado de lineamientos sobre qué aspectos mencionar. Por tanto, asociado a la tarea se incluyen criterios requeridos (por ejemplo, “Where did you go?”) y opcionales (por ejemplo, “What did you like to do?”) que el estudiante debe abordar en su respuesta. Este tipo de tarea permite evaluar la capacidad del estudiante para producir un discurso más extenso y estructurado.
- **Tipo 4 - Tarea de Comparación y Contraste (1 tarea):** Ejercicios que implican la identificación y descripción de diferencias entre dos imágenes. Este tipo de tarea evalúa habilidades de observación y descripción detalladas.

Notar entonces que se tienen un total de 11 imágenes diferentes utilizadas en las tareas con contexto de referencia (9 en tareas de tipo 1, y 2 en la tarea de tipo 4). Es interesante destacar que para las preguntas de tipo 2 se generaron descripciones textuales asociadas a las imágenes, que sirven como contexto de referencia ya que, como veremos en la Sección 4.6, la herramienta construida no tiene capacidad de visión integrada. Mientras que para la tarea de tipo 4 se proporciona una lista de diferencias como referencia para evaluar las respuestas.

Inicialmente, para cada una de estas tareas se asociaron respuestas obtenidas de **dos fuentes**. Por un lado, se utilizaron **respuestas auténticas de estudiantes**, provenientes de la base de datos de Ceibal, que proporcionaron una visión real de las habilidades y desafíos de los estudiantes. Por otro lado, se incluyeron **respuestas generadas específicamente**, con el objetivo de complementar la información existente.

Posteriormente, para enriquecer el conjunto de datos y aumentar la cantidad de respuestas, se generaron datos sintéticos utilizando modelos de lenguaje avanzados como **GPT-4** y **Claude 3.5 Sonnet**. Este proceso consistió en proporcionar ejemplos específicos y solicitudes estructuradas, seleccionando posteriormente las respuestas más relevantes y de mayor interés para el estudio. Estas **respuestas adicionales** fueron revisadas y validadas para ser incluidas en el conjunto de datos.

Es importante notar que las respuestas generadas de esta manera no tienen **audios asociados**, a diferencia de las respuestas auténticas. Sin embargo, como se detalla más adelante en la sección 4.4, esta limitación no representa un obstáculo significativo ya que nuestra propuesta es **modular**, lo cual permite trabajar con transcripciones de los audios, manteniendo la utilidad del análisis.

Etiquetado de la base de datos: Todas las respuestas de la base fueron etiquetadas en términos de task fulfillment siguiendo tres categorías:

- **Not Addressed (No Abordada):** Respuestas que no responden adecuadamente a la tarea planteada, considerándose incorrectas o irrelevantes.
- **Partially Addressed (Parcialmente Abordada):** Respuestas que abordan parcialmente la tarea, pero pueden contener errores o no cubrir todos los aspectos requeridos.
- **Fully Addressed (Completamente Abordada):** Respuestas que comunican de manera efectiva una solución completa a la tarea, independientemente de errores gramaticales o variaciones en el uso del lenguaje que no afecten la comprensión del mensaje.

Esta clasificación se implementó considerando que la categoría *Partially Addressed* abarca respuestas que, aunque contienen elementos correctos, no son completamente precisas, permitiendo así una diferenciación clara entre respuestas correctas e incorrectas.

No se agregaron más matices en el etiquetado debido al alto consumo de tiempo que implicaría construir un conjunto de datos con características más detalladas, que sea consistente. Además, para ciertos tipos de tareas, especialmente los tipos 1 y 2, estos matices no parecían necesarios dada la naturaleza de las preguntas y lo que se espera de las respuestas.

Es importante mencionar que todos los datos fueron etiquetados manualmente por nosotros, de modo de lograr uniformidad en criterio en la clasificación de cada respuesta.

Estadísticas de la Base de Datos:

La base de datos incluye un total de 452 respuestas, distribuidas entre los cuatro tipos de pregunta y de fuente de datos como se detalla en la Tabla 4.1.

Tipo de Tarea	Ceibal	Custom*	GPT-4	Claude 3.5 Sonnet	Total
1	140	41	10	96	287
2	49	31	16	0	96
3	8	3	7	37	55
4	2	14	0	0	16
Total	199	89	33	133	454

Tabla 4.1: Cantidad de respuestas por fuente y tipo de tarea
*Custom se refiere a respuestas generadas manualmente por nosotros

Para cada tipo de pregunta, la distribución de la etiqueta entre sus tres valores posibles se ve de la Tabla 4.2.

Tipo de Tarea	Fully Addressed	Partially Addressed	Not Addressed	Total
1	117	75	95	287
2	42	17	37	96
3	26	18	11	55
4	8	4	4	16
Total	193	114	147	454

Tabla 4.2: Distribución de etiquetas por tipo de tarea

Esta base de datos es un recurso de utilidad para desarrollar y evaluar nuestro sistema de evaluación automatizada del habla.

4.4. Arquitectura del Sistema de Evaluación Propuesto

Para abordar el desafío de la evaluación automatizada del *task fulfillment*, se pensó en un sistema modular que integra diferentes tecnologías de procesamiento de lenguaje natural. Esta sección presenta una visión general de la arquitectura del sistema, sus componentes principales y cómo se adapta específicamente a la tarea de evaluación del cumplimiento de tareas en el contexto de pruebas de habla en inglés.

4.4.1. Visión General del Sistema

Nuestro sistema de evaluación automatizada está diseñado como una serie de módulos interconectados, cada uno especializado en una tarea específica dentro del proceso de evaluación. Esta arquitectura modular facilita la integración de

nuevas funcionalidades y la adaptación del sistema a futuras necesidades de evaluación.

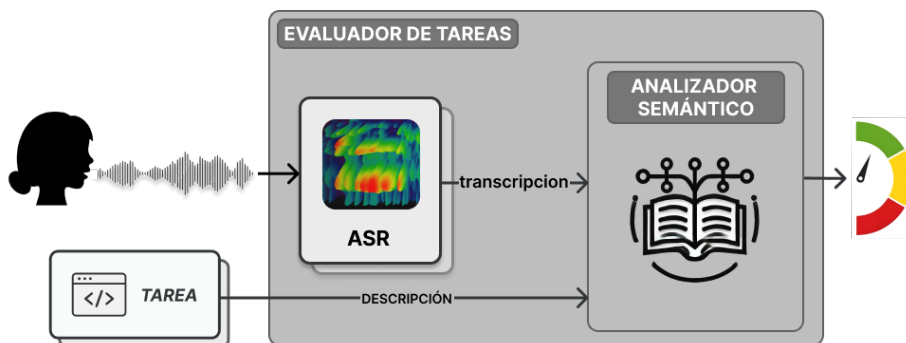


Figura 4.1: Esquema general del sistema de evaluación de tareas.

Como se puede ver en la Figura 4.1, el sistema se compone de dos módulos principales, cada uno diseñado para abordar una parte del proceso de evaluación:

Módulo de Reconocimiento Automático del Habla

Este módulo es responsable de la transcripción del audio de las respuestas orales de los estudiantes a texto, utilizando un sistema de reconocimiento automático del habla para procesar y convertir el lenguaje hablado en su forma escrita. Dada la naturaleza de nuestro público objetivo, este módulo debe ser suficientemente robusto para manejar las características del habla no nativa. Esto incluye la robustez a variaciones en la pronunciación, conteniendo distintos acentos, entre otras.

Además, si bien se asume que los audios no tendrán ruido significativo y/o variaciones notorias en calidad, la robustez frente a estos aspectos es un aspecto de interés para una uniformidad en la evaluación, por lo tanto se hará una mención al respecto al seleccionar el módulo encargado de transcribir.

Módulo de Analizador Semántico

Este es el módulo central del sistema y se encarga de analizar el contenido semántico de las respuestas transcritas. Emplea técnicas de procesamiento de lenguaje natural para evaluar el cumplimiento de la tarea (task fulfillment), considerando los diferentes tipos de tareas identificados en la Sección 2.3.

Entre las características clave que se considera que el analizador semántico debería poseer se destacan: una sólida comprensión contextual que permita interpretar las respuestas en el marco específico de cada tarea, robustez lingüística para manejar variaciones típicas de estudiantes de inglés como segundo idioma (incluyendo errores gramaticales y de vocabulario), capacidad para realizar una evaluación matizada que distinga entre diferentes niveles de cumplimiento de la

tarea, y la habilidad para proporcionar una justificación clara de su evaluación, aumentando así la transparencia del proceso.

Flujo de trabajo del sistema

El proceso de evaluación sigue el siguiente flujo:

1. El audio de la respuesta del estudiante se ingresa al módulo ASR para su transcripción.
2. La transcripción resultante se pasa al módulo de Analizador Semántico junto con la información específica de la tarea.
3. El Analizador Semántico procesa la transcripción en el contexto de la tarea asignada y genera una evaluación del task fulfillment, potencialmente retornando mas información de interés.

La separación de estos componentes permite la evaluación y optimización independiente de cada uno, lo cual se aprovechó para realizar un análisis del *task fulfillment* en dos fases distintas pero complementarias. En las siguientes secciones, se profundiza en cada uno de los componentes principales de este sistema. Primero, se explora el módulo de Reconocimiento Automático del Habla, detallando su importancia, las opciones consideradas y la selección final de acuerdo con criterios definidos. Posteriormente, se aborda el Analizador Semántico, examinando los fundamentos de los Modelos de Lenguaje de Gran Escala (LLMs) y su aplicación en este sistema de evaluación.

Una de las ventajas clave de este enfoque es la posibilidad de trabajar de manera aislada con el componente de análisis semántico, permitiendo el uso de datos sintéticos además de transcripciones de audios reales. Esta estrategia acelera el proceso de iteración y prueba, especialmente considerando las limitaciones de recursos computacionales disponibles.

A continuación, se presentan en detalle cada uno de los componentes mencionados, comenzando con el módulo de Reconocimiento Automático del Habla.

4.5. Reconocimiento Automático del Habla

El Reconocimiento Automático del Habla es un componente esencial en el sistema de evaluación automatizada. Su función principal es convertir las respuestas orales de los estudiantes en texto, proporcionando la base para el análisis semántico posterior. En esta sección, se describe el proceso de selección y evaluación del sistema ASR más adecuado para la tarea específica.

Notar que la precisión del ASR es fundamental para garantizar una evaluación justa y precisa del task fulfillment. Un sistema ASR robusto debe ser capaz de manejar las particularidades del habla no nativa, incluyendo acentos variados, pronunciaciones imprecisas y posibles errores gramaticales, sin perder el contenido semántico de las respuestas.

4.5.1. Comparativa de Sistemas ASR

Para seleccionar el sistema ASR más adecuado, se evaluaron cuatro sistemas diferentes:

- **Google Speech-to-Text:** Un servicio de ASR de propósito general desarrollado por Google y con capacidad para manejar múltiples idiomas y acentos [49].
- **Whisper:** Un modelo de ASR de código abierto desarrollado por OpenAI, que ha demostrado un rendimiento robusto en diversos escenarios, incluyendo habla no nativa [50]. Está basado en la arquitectura de Transformer [51]. Se usará en su versión “large”.
- **WhisperX:** Una versión mejorada de Whisper que incorpora optimizaciones adicionales para mayor velocidad y precisión, que además cuenta con características interesantes como el uso de un VAD (Voice activity detection), diarización, entre otras [52]. Este modelo se usará en su versión “large”.
- **Wav2Vec 2.0 for CTC:** Una adaptación del modelo Wav2Vec 2.0 para ASR que utiliza Connectionist Temporal Classification (CTC). Este algoritmo permite mapear secuencias de audio a texto sin necesidad de un alineamiento explícito, combinando el aprendizaje auto-supervisado con una alineación eficiente [53]. Se utilizaron los pesos de “facebook/wav2vec2-large-960h” en hugging face.

Para la evaluación utilizamos la base de datos EpaDB, presentada en la sección 3.2.1. Esta base es especialmente útil para el análisis de los sistemas ASR, ya que cada audio cuenta con su texto de referencia asociado. En particular, al igual que se hizo para la evaluación de características acústicas se considera el *EpaDB subset* presentado anteriormente.

El proceso de evaluación consiste en, por cada ASR, procesar todos los audios seleccionados y luego comparar las transcripciones obtenidas contra las transcripciones de referencia a través de las siguientes métricas:

- **Word Error Rate (WER):** Mide la distancia de edición entre la transcripción del ASR y la referencia manual, normalizada por la longitud de la referencia.
- **Character Error Rate (CER):** Similar al WER, pero a nivel de caracteres.

Además se medirá el **tiempo de procesamiento promedio** sobre todo el conjunto de datos.

4.5.2. Resultados y Selección del ASR

Los resultados promedio obtenidos para cada sistema ASR en el conjunto de datos considerado se presentan en la Tabla 4.3.

ASR	WER	CER	Tiempo
Google	0.1151	0.0639	0.7341
Wav2Vec2	0.2593	0.1147	0.0265
Whisper	0.0841	0.0428	0.9257
WhisperX	0.0813	0.0407	0.3041

Tabla 4.3: Resultados globales de los sistemas ASR

Como se puede observar, estos muestran que WhisperX tiene el mejor rendimiento en términos de WER y CER, seguido de cerca por Whisper. Google Speech-to-Text muestra un rendimiento medio, mientras que Wav2Vec2 presenta el mayor error en ambas métricas. En cuanto al tiempo de procesamiento, Wav2Vec2 es el que menos demora, seguido por WhisperX, aunque esto debe considerarse en relación a la precisión.

Análisis por Hablante

El análisis por hablante muestra variaciones significativas en el rendimiento de los sistemas ASR. La Figura 4.2 muestra el WER promedio para cada sistema ASR por hablante.

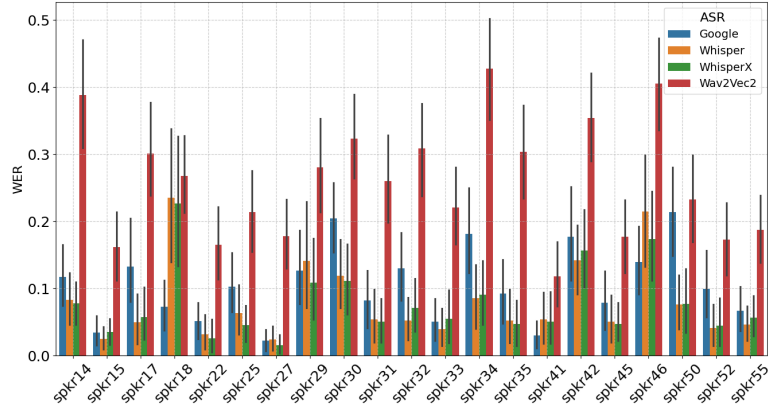


Figura 4.2: WER promedio por hablante para cada sistema ASR

Se observa que WhisperX y Whisper mantienen consistentemente un mejor rendimiento promedio en todos los hablantes, mientras que Wav2Vec2 muestra

la mayor variabilidad y los valores más altos de WER. El de Google, logra resultados intermedios, siendo peores a los WhisperX y Whisper, y mejores a Wav2Vec2.

Análisis por Número de Audio

El análisis por número de audio permite identificar patrones en el rendimiento de los ASR según el contenido específico de cada grabación. Recordar que para igual número de audio, los hablantes dicen la misma frase. Por tanto, agrupando por número de audio y promediando se obtiene el promedio de WER de los ASR para una frase específica sobre los 21 hablantes. La Figura 4.3 muestra este WER promediado por audio para cada sistema ASR.

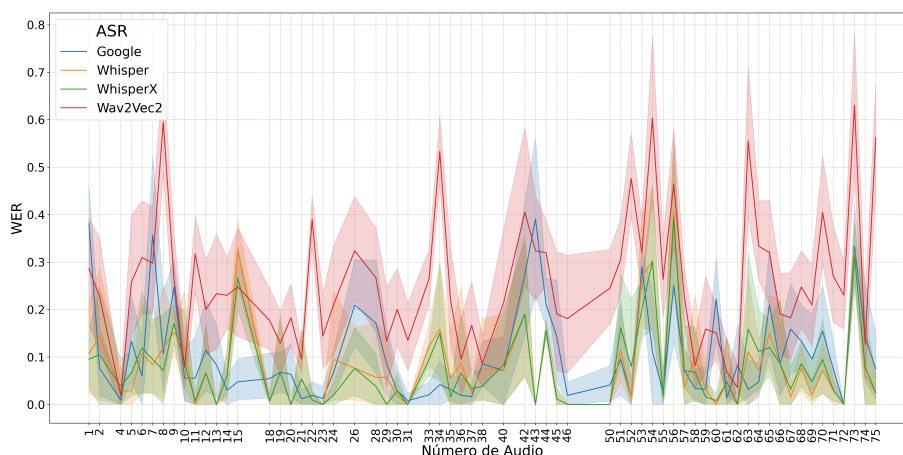


Figura 4.3: WER por número de audio para cada sistema ASR

Este análisis deja notar que ciertos audios son particularmente desafiantes para todos los sistemas, posiblemente debido a características específicas del contenido. WhisperX y Whisper muestran una mayor robustez ante estas variaciones.

Selección del Sistema ASR y Justificación

Con base en los resultados obtenidos y el conocimiento sobre estos modelos, se seleccionó WhisperX como el sistema ASR más adecuado para la tarea de evaluación. Esta decisión se justifica en primer lugar porque WhisperX demostró una precisión superior, exhibiendo el mejor rendimiento tanto en WER (Word Error Rate) como en CER (Character Error Rate). Este nivel de precisión es necesario para asegurar un desempeño óptimo del modelo, incluso al evaluar hablantes con diferentes niveles de competencia en inglés. Además, WhisperX se distinguió por una velocidad de procesamiento aproximadamente 3 veces mayor

que la de Whisper, a menor WER y CER. Si bien el transcriptor basado en Wav2Vec2 fue el más rápido, su WER y CER no fueron buenos en comparación.

Otro aspecto destacable de WhisperX es su robustez ante la variabilidad lingüística. Durante las pruebas, el sistema mostró un rendimiento más consistente a través de diferentes hablantes, lo que sugiere una mayor adaptabilidad a diversas condiciones del habla y el ruido. Además, el código abierto de WhisperX permite adaptaciones y personalizaciones futuras según las necesidades del proyecto. Su precisión, velocidad, robustez y flexibilidad lo convierten en una opción adecuada para la tarea de evaluación de cumplimiento.

4.6. Analizador Semántico: Fundamentos y Diseño

Habiendo ya explorado el módulo de Reconocimiento Automático del Habla, ahora nos centramos en el Analizador Semántico. Como ya se mencionó en 4.4.1, este módulo es el componente central de nuestro sistema de evaluación automatizada, responsable de interpretar y evaluar el contenido semántico de las respuestas transcritas, determinando el *task fulfillment*.

En esta sección, examinaremos la estructura interna del Analizador Semántico, los fundamentos teóricos que sustentan su diseño, y los aspectos específicos de su implementación. Comenzaremos presentando la arquitectura interna del módulo, desglosando sus principales subcomponentes y sus interacciones. Luego, profundizaremos en los Modelos de Lenguaje de Gran Escala (LLMs), los cuales son parte fundamental de la propuesta planteada. Finalmente, abordaremos los detalles de diseño e implementación.

4.6.1. Arquitectura Interna del Analizador Semántico

El Analizador Semántico está diseñado como un sistema modular compuesto por tres subcomponentes principales: el Constructor de Prompts, el Modelo de Lenguaje de Gran Escala (LLM) y el Parser de Resultados. Esta estructura modular ofrece flexibilidad en el diseño y facilita la posibilidad de modificar o actualizar componentes individuales según sea necesario.

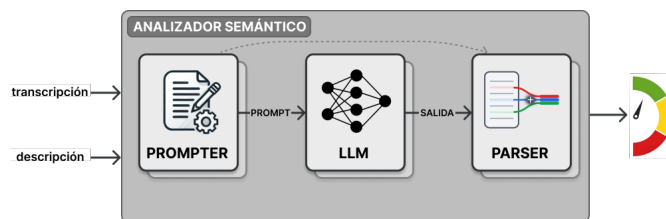


Figura 4.4: Arquitectura del Analizador Semántico.

La Figura 4.4 muestra un diagrama de flujo que ilustra la arquitectura y el

proceso del Analizador Semántico. Cada componente cumple un rol específico como se describe a continuación:

En primer lugar, el **Constructor de Prompts** recibe la información de la tarea y la respuesta transcrita. Utiliza un conjunto de plantillas pre-diseñados, específicos para cada tipo de tarea, para generar un prompt que incorpora los detalles de la tarea, la respuesta del estudiante y las directrices para la evaluación.

Acto seguido, el **Modelo de Lenguaje** procesa el prompt generado para realizar la evaluación semántica.

Finalmente, el **Parser de Resultados** interpreta la salida del modelo de lenguaje y extrae una evaluación estructurada del task fulfillment.

Con la arquitectura del Analizador Semántico definida, se procede a presentar los fundamentos y la aplicación de los Modelos de Lenguaje de Gran Escala (LLMs) en las tareas de evaluación. Esta subsección expone el uso de los LLMs para la evaluación y destaca la selección del modelo más adecuado para el estudio.

4.6.2. Modelos de Lenguaje de Gran Escala (LLMs) en Evaluación Semántica

Los Modelos de Lenguaje de Gran Escala (LLMs, por sus siglas en inglés) han transformado el campo del procesamiento del lenguaje natural y desempeñan un papel esencial en el enfoque de análisis semántico. Estos modelos de aprendizaje profundo, basados en arquitecturas neuronales avanzadas, están diseñados para comprender y generar lenguaje natural de manera eficiente. Se distinguen por su gran escala, reflejada tanto en el elevado número de parámetros como en la abundante cantidad de datos utilizados durante su entrenamiento.

La evolución de los LLMs ha sido rápida y transformadora, estableciendo hitos en términos de capacidad de modelado y comprensión del lenguaje natural. El concepto de los LLMs, tal como los conocemos hoy en día, surgió con la introducción de la arquitectura Transformer por [51]. Esta arquitectura, basada enteramente en mecanismos de atención, permitió el entrenamiento eficiente de modelos mucho más grandes que sus predecesores. A partir de este avance, una sucesión de modelos cada vez más potentes ha redefinido los límites del procesamiento del lenguaje natural, incluyendo BERT [54], GPT [55], GPT-3 [56], y más recientemente GPT-4 [57] y LLaMA 3 [58], los cuales ofrecen una comprensión aún más profunda y precisa del lenguaje.

El potencial de los LLMs va más allá de la generación de texto; su capacidad para comprender el significado y las relaciones entre distintos componentes del lenguaje ha facilitado su aplicación en diversas áreas, incluida la evaluación semántica. Por ello, resulta pertinente explorar cómo estas capacidades avanzadas de comprensión del lenguaje pueden ser aprovechadas específicamente en el ámbito de la evaluación educativa.

LLMs en evaluación educativa

En el contexto de la evaluación educativa, los Modelos de Lenguaje de Gran Escala representan un avance significativo en la automatización de la enseñanza, ya que permiten un análisis detallado y contextual de las respuestas de texto libre. Estos modelos destacan por su capacidad para procesar y evaluar el lenguaje de manera integral, lo que los hace especialmente valiosos en el ámbito educativo.

Una de las principales ventajas de utilizar LLMs en tareas de evaluación, en comparación con otras técnicas de procesamiento de lenguaje, es su capacidad para entender el contexto y las sutilezas del lenguaje. Esto permite una evaluación más holística, capaz de identificar errores gramaticales, contextuales y de coherencia, yendo más allá de la coincidencia de palabras clave. Por ejemplo, en el estudio realizado por Taghipour et al., se demostró que los LLMs pueden evaluar eficazmente la calidad de ensayos cortos, mostrando una alta correlación con las evaluaciones humanas [59].

Otra aplicación prometedora es el uso de LLMs para proporcionar retroalimentación detallada y personalizada. Nguyen et al. exploraron cómo los LLMs pueden generar explicaciones y sugerencias de mejora para respuestas de estudiantes, complementando así la labor de los docentes [60].

Sin embargo, el uso de LLMs en la evaluación educativa también presenta desafíos. La interpretabilidad de las decisiones del modelo sigue siendo un área de preocupación, ya que es importante entender cómo y por qué un LLM llega a una determinada evaluación. Además, los sesgos inherentes en los datos de entrenamiento pueden llevar a evaluaciones injustas o inexactas, especialmente cuando se trata de estudiantes de diversos orígenes lingüísticos y culturales.

Técnicas Principales de Uso de LLMs

Existen diversas técnicas para aplicar los LLMs en tareas específicas como la evaluación semántica. Las principales son:

La primer técnica, conocida como **Zero-shot Learning** o **Aprendizaje sin ejemplos**, consiste en utilizar al modelo para realizar una tarea sin proporcionarle ejemplos concretos ni entrenamiento adicional. En el contexto de la evaluación de *task fulfillment*, esto podría significar solicitar al modelo que evalúe una respuesta basándose únicamente en la pregunta y en criterios generales. La principal ventaja de esta técnica radica en su flexibilidad, ya que permite al modelo abordar tareas nuevas sin necesidad de preparación previa. No obstante, es necesario redactar instrucciones claras y precisas que permitan al modelo comprender la tarea deseada.

Por otro lado, a diferencia de la primera técnica, se encuentra el **Few-shot Learning** o **Aprendizaje con pocos ejemplos**. En este enfoque, se proporcionan al modelo algunos ejemplos de la tarea deseada dentro del prompt. Esto puede mejorar significativamente el rendimiento en comparación con el aprendizaje sin ejemplos, especialmente para tareas específicas o poco comunes. Brown et al. demostraron la eficacia de esta técnica en una variedad de tareas

lingüísticas [56]. En este contexto, se pueden proporcionar al modelo ejemplos de respuestas correctas e incorrectas antes de solicitarle que evalúe una nueva respuesta, lo que podría ayudar al modelo a comprender mejor los criterios de evaluación específicos. Cabe señalar que, idealmente, se deben elegir ejemplos representativos y presentarlos de manera que el modelo pueda generalizar adecuadamente.

Fine-tuning (Ajuste fino del modelo) El *fine-tuning* consiste en ajustar un modelo preentrenado utilizando un conjunto de datos específico de la tarea. Aunque este proceso requiere más recursos y una cantidad considerable de datos etiquetados, puede mejorar significativamente el rendimiento en tareas concretas. En el ámbito de la evaluación educativa, se puede aplicar para personalizar un LLM según los estándares de una institución o programa, lo que puede ayudar a que el modelo comprenda y utilice de manera más adecuada los criterios de evaluación establecidos.

En todas estas técnicas, el diseño efectivo del prompt y la interacción cuidadosa con el modelo son fundamentales para lograr resultados óptimos en la evaluación de *task fulfillment*.

Selección del Modelo para el Proyecto

Actualmente, el panorama de los Modelos de Lenguaje de Gran Escala es diverso y ofrece múltiples opciones adaptables a las necesidades de un sistema de evaluación. Entre los más notables se encuentra *GPT-4*, desarrollado por OpenAI, que representa la vanguardia en modelos comerciales de gran escala [61]. Google, por su parte, ha introducido la familia *Gemini*, que se distingue por sus capacidades multimodales [62]. Anthropic ha contribuido con *Claude Sonnet*, parte de su línea *Claude 3*, que se ha posicionado como una alternativa competitiva en el mercado de LLMs comerciales [63].

En contraste con estos modelos propietarios, Meta ha lanzado *LLaMA 3*, un modelo de código abierto que ha ganado considerable atención en la comunidad de investigación. La disponibilidad de sus pesos y su rendimiento competitivo lo han convertido en una opción atractiva para aplicaciones académicas y comerciales [64].

Para el presente proyecto de evaluación automatizada del habla en inglés, se seleccionó el modelo **LLaMA 3 de 8 mil millones de parámetros**. Esta elección se justifica principalmente a partir de las siguientes consideraciones:

En primer lugar, **la accesibilidad y licencia**: los pesos del modelo son accesibles bajo una licencia que permite su uso en investigación y aplicaciones como la presente. Esto proporciona la flexibilidad necesaria para posibles adaptaciones futuras, además de permitir su ejecución de manera local en hardware dedicado.

En segundo lugar, **la eficiencia computacional**: con 8 mil millones de parámetros, este modelo ofrece un buen equilibrio entre capacidad y uso de recursos, permitiendo su ejecución en hardware más accesible.

En tercer lugar, **el rendimiento adecuado**: estudios preliminares y pruebas

realizadas han demostrado que el modelo presenta un desempeño satisfactorio para las tareas de evaluación abordadas.

Además, la posibilidad de ejecutar el modelo localmente resulta fundamental, dado que se trabaja con datos internos relacionados con los estudiantes. Esta implementación permite mantener un mayor control sobre el manejo de la información y se alinea con las buenas prácticas de protección de datos.

Por otro lado, aunque este modelo no presenta capacidades multimodales, esto no representa un obstáculo significativo para el proyecto. Las únicas tareas donde se requerirían dichas capacidades son las de *tipo 2* y *tipo 4*, que involucran imágenes. Para ambos casos, se optó por utilizar descripciones textuales de referencia que contienen la información necesaria para la evaluación, suplantando así la necesidad de capacidades multimodales. Los detalles específicos de esta implementación se abordarán en la Sección 4.6.3.

4.6.3. Diseño e Implementación del Analizador Semántico

Habiendo seleccionado el modelo de lenguaje a utilizar, **LLaMA 3 de 8 mil millones de parámetros**, se opta por emplear el modelo preentrenado en su versión “Instruct”, que es un *fine-tuning* del modelo base para seguir instrucciones y mantener conversaciones en un formato de roles (System, User, Assistant). Esta versión está específicamente ajustada para comprender y ejecutar instrucciones directas, lo que facilita darle órdenes específicas y obtener respuestas coherentes con el rol asignado. En este formato, el rol “System” define el comportamiento y contexto general del modelo, el “User” proporciona las instrucciones o consultas, y el “Assistant” responde siguiendo las pautas establecidas por el “System” y las solicitudes del “User”. No se realizó ningún *fine-tuning* adicional por nuestra parte.

En su lugar, se utilizarán diferentes técnicas de *prompting*, explorando tanto enfoques de **zero-shot** como de **few-shot**. El objetivo es desarrollar y evaluar múltiples templates para determinar cuál de ellos ofrece el mejor desempeño en la evaluación de *task fulfillment* sobre el conjunto de datos anteriormente presentado.

A continuación presentamos las técnicas utilizadas en los *prompts* generados.

Aspectos generales del Diseño de *Prompts*

El diseño efectivo de *prompts* es fundamental para guiar a los modelos de lenguaje en tareas específicas, como la evaluación de respuestas en un contexto educativo. En este proyecto, se han explorado diversas técnicas de *prompting* para intentar maximizar el desempeño del modelo **LLaMA 3** en la tarea de evaluación automatizada del *task fulfillment*.

Las técnicas generales de *prompting* utilizadas incluyen:

- **Zero-shot prompting:** Consiste en proporcionar al modelo una instrucción directa sin ejemplos adicionales. Se espera que el modelo genere una

respuesta basada únicamente en su conocimiento previo y la instrucción dada.

- ***Few-shot prompting***: Implica proporcionar al modelo algunos ejemplos de entradas y salidas deseadas antes de solicitar una respuesta a una nueva entrada. Esto ayuda al modelo a entender mejor el formato y el tipo de respuesta esperada [65].
- ***Chain-of-thought prompting***: Esta técnica guía al modelo a través de un proceso de razonamiento paso a paso antes de llegar a una conclusión final. Puede ser *libre* (sin estructura específica) o *guiado* (con pasos estructurados) [66].
- **Evaluación categórica y escalas simplificadas**: Se utilizan categorías predefinidas o escalas numéricas simples para facilitar la evaluación y clasificación de las respuestas.
- **Técnicas de rol o simulación**: Se instruye al modelo para que asuma un rol específico, como el de un docente empático, para evaluar las respuestas desde una perspectiva particular.

Estas técnicas se seleccionaron y adaptaron en función de las necesidades específicas de cada tipo de tarea, buscando obtener los mejores resultados que fueran posibles. La construcción de los *prompts* se hizo de manera experimental, inicialmente generando algunos sencillos y viendo cómo se comportaba el modelo ante estos, luego agregando más detalle o aclaraciones en caso de que fuera necesario para intentar resolver problemas encontrados.

A continuación se detalla cómo fueron diseñados los *prompts* para cada tipo de tarea presentada en la sección 4.3, utilizando las técnicas mencionadas anteriormente.

Como comentario general a todos, se destaca el uso del **formato JSON** como respuesta esperada al realizar inferencia sobre el modelo, solicitud que el modelo respeta de manera adecuada en la mayoría de los casos, según los experimentos realizados. Gracias a este formato, el parser de resultados logra interpretar la salida del modelo de lenguaje y estructurarla de manera adecuada para su evaluación final, utilizando expresiones regulares para convertir estas respuestas en objetos manejables. Luego, asigna las salidas relevantes a las categorías definidas mediante un mapeo asociado a cada *prompt*. Por ejemplo, se realiza el mapeo de una escala numérica a las categorías preestablecidas.

Además, todos los *prompts* fueron creados en idioma inglés. Esto dado que las consignas y respuestas se encuentran en dicho idioma, por lo tanto, el uso del mismo en todo el *prompt* garantiza coherencia lingüística y minimiza ambigüedades de traducción.

Para mantener el enfoque del documento, se incluyen únicamente algunos *prompts* representativos como ejemplo. Los restantes se encuentran detallados en el anexo (Véase A.4).

Diseño de *Prompts* para Tareas Tipo 1

Para las tareas de Tipo 1, se desarrollaron múltiples *prompts* con el objetivo de evaluar diversas estrategias para calificar la relevancia de la respuesta del estudiante con respecto a la pregunta planteada. Solo se definió como entrada la pregunta y la respuesta dada por el estudiante. A continuación, se presentan algunos de los generados. La lista completa de *prompts* se encuentra en el Anexo [A.4.1](#).

■ Evaluación paso a paso

Descripción: Implementa *chain of thought*, solicitando al modelo que, antes de determinar la relevancia de una respuesta, siga un proceso de pensamiento de tres pasos, considerando que respuestas breves pueden ser igualmente válidas.

```
SYSTEM: You are a supportive English teacher with
        expertise in evaluating spoken language
        learners' responses to open-ended questions.
```

```
USER: You will need to rate a student's answer
      based on its relevance to the question.
      Assess the relevance of the given answer to
      the question by breaking down your thought
      process step-by-step. Provide your assessment
      in the following JSON format: {"
      thought_process": ["step1", "step2", "step3
      "], "relevance": "category"}.
```

```
Use one of these categories: "Not Addressed", "
      Partially Addressed", or "Fully Addressed".
```

```
Keeping in mind that brief responses can be fully
      valid.
```

```
Question: {question}
Student's answer: {answer}
Assessment:
```

■ Evaluación con 3 ejemplos

Descripción: Implementa *few-shot prompting* sobre un *prompt* prácticamente sin detalle, proporcionando al modelo tres ejemplos concretos de preguntas, respuestas y evaluación esperada. Se espera que el modelo infiera parte de los criterios de evaluación.

```
SYSTEM: You are an expert English language
        evaluator specialized in assessing spoken
        responses to simple, open-ended questions.
```

USER: Evaluate if the student's answer is relevant to the given question. Provide your assessment ONLY in the following JSON format: {relevance: category, justification: brief explanation} Use one of these categories for relevance: Not Addressed, Partially Addressed, or Fully Addressed."

Here are some examples:

Question: Can you swim?
Student's answer: Yes I can
Assessment: {"relevance": "Fully Addressed", "justification": "Direct and relevant answer to the question."}

Question: What is your favourite breakfast?
Student's answer: I like swimming
Assessment: {"relevance": "Not Addressed", "justification": "Answer is not related to breakfast preferences."}

Question: What time do you go to school?
Student's answer: In the morning
Assessment: {"relevance": "Partially Addressed", "justification": "Provides a general timeframe but not a specific time."}

Now, evaluate the following:
Question: {question}
Student's answer: {answer}
Assessment:

Diseño de *Prompts* para Tareas Tipo 2

Para las tareas de Tipo 2, que involucran preguntas basadas en imágenes, y para abordar las limitaciones inherentes de los modelos sin capacidad de visión, todos los *prompts* se diseñaron proporcionando al modelo una **descripción detallada de la imagen** como contexto para evaluar la correctitud de las respuestas. Esta descripción no representa una respuesta modelo, sino que establece el marco de referencia para validar los elementos centrales y reconocer detalles adicionales mencionados por el estudiante.

Esta simplificación es efectiva siempre y cuando la pregunta sea específica y la descripción de la imagen sea lo suficientemente detallada para contener la

información necesaria de la respuesta. De lo contrario, podría suceder que el alumno mencione aspectos que, aunque correctos y presentes en la imagen, no estén incluidos en la descripción proporcionada, lo que podría afectar la precisión de la evaluación.

La evaluación prioriza la **precisión en la identificación del elemento central** de la pregunta por sobre los detalles proporcionados, aunque estos últimos son igualmente contrastados con la descripción de referencia para verificar su validez. Por ejemplo, ante la pregunta “What vehicle can you see?” y una descripción “old red motorcycle in deteriorated condition”, la respuesta “A motorcycle” recibiría la clasificación “**Fully Addressed**” por identificar correctamente el vehículo en cuestión. En contraste, una respuesta como “A green motorcycle” se clasificaría como “**Partially Addressed**” debido a que, si bien identifica el tipo de vehículo, incluye un detalle que contradice la descripción de referencia. Una respuesta como “A red bike” se clasificaría como “**Not Addressed**”, ya que, a pesar de mencionar el color correcto, falla en identificar apropiadamente el tipo de vehículo principal.

Se diseñaron diferentes variantes de *prompts* empleando **tres estrategias principales**: *explicación de criterios*, *chain of thought*, y *few-shot learning*. El detalle completo de estos *prompts* se encuentra en el Anexo [A.4.2](#).

Como parte del formato de salida, para algunos *prompts* se incorporaron elementos adicionales más allá de la mera categorización, incluyendo una **lista de detalles adicionales** mencionados por el estudiante que no son contrastables con la descripción de referencia. Este último elemento pretende ayudar al modelo a diferenciar entre aspectos evaluables y no evaluables, a la vez que proporcionar retroalimentación valiosa sobre elementos que, estando potencialmente presentes en la imagen, podrían no haber sido capturados en la descripción textual proporcionada.

Diseño de *Prompts* para Tareas Tipo 3

Para las tareas de Tipo 3, que involucran respuestas elaboradas con criterios requeridos específicos, los *prompts* implementaron una estrategia de evaluación que contempla dos aspectos fundamentales: la **adherencia al tema principal** y el **cumplimiento de los criterios solicitados**.

Por ejemplo, si la tarea solicita describir una experiencia personal y establece tres criterios específicos de desarrollo, una respuesta clasificada como “**Fully Addressed**” debe mantener la coherencia temática y abordar todos los puntos solicitados. En contraste, una respuesta que mantenga el tema pero omita algunos criterios recibiría una clasificación “**Partially Addressed**”. En todos los *prompts* se incorporó, como parte de la salida esperada, una **lista** detallada de los **criterios cubiertos y omitidos**, lo que no solo facilita la validación posterior de la evaluación sino que también sirve como guía para que el modelo determine la clasificación final.

Se diseñaron diferentes variantes de *prompts* empleando dos estrategias principales: *explicación de criterios* y *chain of thought*. El detalle completo de estos *prompts* se encuentra en el Anexo [A.4.3](#).

Adicionalmente, se exploró un enfoque alternativo basado en una **evaluación por etapas**, donde se segmenta el proceso en múltiples consultas secuenciales dentro de una misma conversación: primero se evalúa la **adherencia al tema principal**, luego, manteniendo el contexto de la evaluación anterior, se analiza el cumplimiento de los requerimientos clasificándolos en criterios explícitamente mencionados, implícitamente abordados o no contenidos, y finalmente, aprovechando todo el análisis previo, se integran estos resultados para determinar la clasificación final de *task fulfillment*. Este enfoque conversacional permite que cada etapa se beneficie del razonamiento y las conclusiones de las etapas anteriores, evitando la necesidad de un prompt inicial extenso y complejo, y permitiendo centrarse en aspectos específicos en cada paso del análisis.

Diseño de *Prompts* para Tareas Tipo 4

Las tareas de Tipo 4 requieren que el estudiante identifique una cantidad mínima de diferencias específicas entre dos imágenes. La característica distintiva de estas tareas es que existe un **conjunto finito y predefinido de diferencias válidas**, lo que permite diseñar una evaluación conceptualmente similar a la de las tareas de Tipo 3, donde la lista de diferencias válidas actúa de manera análoga a los criterios requeridos. Adicionalmente, esta abstracción permite simplificar la evaluación sin necesidad de proporcionar descripciones detalladas de las imágenes como se realizó en las tareas de Tipo 2.

Se desarrolló una evaluación integral donde se le proporciona al modelo la lista completa de diferencias válidas y la respuesta del estudiante. El modelo debe generar tres listas resultantes: diferencias acertadas (presentes en la lista de referencia), con justificación de esto (extracto del texto que muestra el acierto), diferencias no identificadas (ausentes en la respuesta dada), y diferencias incorrectas (mencionadas en la respuesta, pero fuera de la lista de referencia). Con base en estas listas generadas, el modelo genera una calificación final considerando tanto la cantidad mínima necesaria de diferencias acertadas como la cantidad de errores presentes.

El detalle completo del *prompt* desarrollado para ambos enfoques se encuentra en el Anexo [A.4.4](#).

4.6.4. Reflexión general sobre el Diseño de *Prompts*

El proceso de diseño de *prompts* fue iterativo y presentó diversos desafíos. A continuación, se presentan algunas reflexiones generales sobre este proceso:

Uno de los principales desafíos fue asegurar que el modelo interpretara correctamente las instrucciones y proporcionara las evaluaciones en el formato deseado. En ocasiones, el modelo tendía a generar texto adicional o desviarse del formato JSON especificado, e incluso dar categorías inventadas. Para mitigar esto, se ajustaron las instrucciones y se reforzó el énfasis en el formato de respuesta requerido, así como en las categorías a considerar. Esto tuvo un

impacto positivo, que se vio reflejado en las evaluaciones realizadas. Por ejemplo, en versiones iniciales de algunos *prompts* con *few shots*, a veces inventaban categorías como “Slightly Addressed” por no especificarse el rango de categorías válidas.

Otro problema interesante surgió al evaluar preguntas de preferencias personales, como “What’s your favourite colour?”. En algunas ocasiones, para algunos *prompt* iniciales, el modelo interpretaba estas preguntas como si estuvieran dirigidas a él mismo como evaluador, lo que llevaba a evaluaciones incorrectas. Por ejemplo, ante la respuesta “My favourite color is blue”, el modelo podía generar evaluaciones como: “The answer is completely irrelevant to the question, as the question asks about the evaluator’s favorite color, but the answer is about the speaker’s favorite color.”

Este comportamiento se corrigió especificando explícitamente en los *prompts* que se trataba de evaluar “student’s answer”, lo que ayudó al modelo a mantener el rol de evaluador sin confundir la perspectiva desde la cual debía analizar las respuestas.

Como comentario extra, los *prompts* más detallados y estructurados pueden conducir a evaluaciones más precisas, pero también incrementan la carga computacional y el tiempo de respuesta del modelo. Por ello, para lograr correr los experimentos en un tiempo razonable, fue importante encontrar un equilibrio entre el nivel de detalle necesario y la eficiencia del sistema.

4.7. Metodología de Evaluación

La evaluación de sistemas automatizados basados en LLMs requiere un enfoque específico debido a su naturaleza estocástica. Estos modelos pueden generar respuestas variadas incluso ante la misma entrada, especialmente cuando se configuran con una temperatura mayor a cero para fomentar diversidad. Esta característica motiva la necesidad de un protocolo de evaluación que considere múltiples ejecuciones. Esta sección describe la metodología empleada para evaluar su rendimiento y confiabilidad.

El **diseño experimental** se basa en la base de datos descrita en la Sección 4.3, que incluye respuestas para los cuatro tipos de tareas identificados. Cabe mencionar que, en los casos con un menor volumen de datos, la representatividad de los resultados cuantitativos puede verse afectada, lo que hace que los análisis cualitativos obtenidos de la evaluación sean más relevantes.

La **construcción de los *prompts*** utilizados en este estudio, tal como se detalló en la Sección 4.6.3, se basó en un enfoque iterativo y empírico, refinando los *prompts* a través de pruebas iniciales y ajustes sucesivos.

La evaluación presentada en esta sección se realiza utilizando la **versión final** de los mismos luego del proceso iterativo.

Destacar que dada la limitada cantidad de datos, utilizamos el conjunto completo tanto para ajustar y probar los prompts como para el análisis final, sin recurrir a particiones dev o test convencionales. Todos los ejemplos, originales y sintéticos, se evaluaron por igual, sin ponderarlos ni extrayendo métricas específicas por cada fuente de datos.

El proceso de evaluación se define entonces de la siguiente manera:

Evaluación por Tipo de Tarea: El primer paso es procesar el conjunto de datos completo a través del sistema para cada tipo de tarea, evaluando cada respuesta con los prompts diseñados. Esta etapa proporciona métricas iniciales de rendimiento como se detalla en la Sección 4.7.1.

Análisis de Consistencia: Dada la naturaleza estocástica de los LLMs, resulta interesante evaluar la estabilidad de las evaluaciones para el análisis del task fulfillment. Por tanto, sería de utilidad realizar múltiples ejecuciones sobre el mismo conjunto de datos, permitiendo:

- Calcular intervalos de confianza para las métricas de evaluación.
- Analizar la variabilidad en las evaluaciones individuales.
- Identificar casos donde el sistema muestra alta variabilidad, si los hubiera.

Se recomienda llevar a cabo un número suficiente de ejecuciones independientes por prompt, típicamente en el rango de 20 a 30, fundamentado en principios estadísticos como la *Ley de los Grandes Números* y el *Teorema del Límite Central*. Realizar este número de ejecuciones permite obtener una distribución representativa de los resultados, facilitando el cálculo de intervalos de confianza estadísticamente significativos para las métricas de evaluación.

Dados los recursos computacionales limitados, se planteó realizar este análisis de consistencia en un único tipo de tarea y sobre los *prompts* de mejor desempeño, en lugar de sobre todos los construidos. Dicho análisis se presenta en la sección 4.8.5.

4.7.1. Métricas

Para evaluar el rendimiento del sistema, se utilizó una combinación de métricas complementarias:

Métricas de Clasificación: Las principales métricas utilizadas incluyen *F1* (media armónica entre precisión y exhaustividad), precisión (proporción de predicciones positivas correctas), y recall o exhaustividad (proporción de casos positivos correctamente identificados). Estas métricas proporcionan una visión general de la capacidad del sistema para clasificar correctamente las respuestas. Notar que, dado que la clasificación se da en 3 categorías, y a su vez la base de datos no se encuentra del todo balanceada, consideramos utilizar estas métricas en su versión ponderada.

Métricas de Acuerdo: Dado que las categorías de evaluación son ordinales (*Not Addressed*, *Partially Addressed*, *Fully Addressed*), utilizamos el Weighted Kappa como la métrica principal de acuerdo. Esta métrica penaliza los desacuerdos basados en la distancia entre categorías, lo que la hace más apropiada que el Cohen’s Kappa estándar.

Interpretamos los valores de Weighted Kappa de la siguiente manera:

- $> 0,8$: Acuerdo muy bueno
- $0,6 - 0,8$: Acuerdo bueno
- $0,4 - 0,6$: Acuerdo moderado
- $< 0,4$: Acuerdo pobre

Análisis de Patrones de Error: Las métricas cuantitativas se complementaron con un análisis cualitativo, utilizando matrices de confusión para visualizar y comprender los patrones específicos de error, así como un análisis de casos de interés, que sirven para mostrar el comportamiento ante diferentes escenarios.

4.8. Resultados y Análisis

En esta sección se presentan los resultados obtenidos a través de la metodología de evaluación empleada y se analizan los mismos. El objetivo es proporcionar una comprensión clara del rendimiento del sistema en función de los distintos tipos de *prompts*, observando sus fortalezas, limitaciones y áreas de mejora potencial.

Los experimentos fueron realizados utilizando el modelo **Llama 3.0 8B**, en su versión quantizada de 8 bits mediante la librería BitsAndBytes. La generación de texto se configuró con **temperatura 0.6** y **top-p 0.9**, parámetros que proporcionan un balance entre consistencia y variabilidad en las respuestas. Se estableció un límite de **2048 tokens** para la generación de nuevas respuestas. Esta configuración permite ejecutar el modelo en hardware de consumo mientras se mantiene un balance adecuado entre rendimiento y uso de recursos. De todas formas, destacar que los tiempos de ejecución fueron considerablemente altos especialmente para *prompts* con bastante detalle.

Los experimentos se ejecutaron en un equipo con las especificaciones que se detallan en la Tabla 4.4.

4.8.1. Análisis para Tareas de Tipo 1

Se presenta aquí el análisis de los resultados obtenidos para las tareas de tipo 1, abordando tanto aspectos cuantitativos como cualitativos. Los *prompts* específicos utilizados se encuentran detallados en el Anexo A.4.1.

Componente	Especificación
GPU	NVIDIA GeForce RTX 4080 Super
CPU	Intel Core i7-14700K @ 3.40 GHz
Memoria	32GB DDR5
Almacenamiento	Samsung 990 PRO SSD

Tabla 4.4: Especificaciones del hardware utilizado

Análisis sobre métricas obtenidas

Los resultados cuantitativos revelan diferencias significativas en el rendimiento de los distintos *prompts* al ser evaluados sobre la base presentada en la Sección 4.3. En la Tabla 4.5 se pueden observar los resultados obtenidos por cada tipo de *prompt*, reportado en las distintas métricas definidas.

Con el propósito de evitar encabezados extensos en las tablas, se recurre a una nomenclatura abreviada que identifica de forma unívoca la estrategia de *prompting* empleada. Los códigos se mantienen de forma consistente en todas las tablas y figuras posteriores; la descripción completa de cada prompt permanece disponible en los apéndices correspondientes.

De esta manera, para las tareas de tipo 1, **cat_simple** designa el prompt base sin instrucciones adicionales; **cat_tol_errores** introduce tolerancia explícita a repeticiones y errores menores; **cat_flex_breve** enfatiza la suficiencia de respuestas breves; **cat_definidas** incorpora criterios de categorización explícitos; **escala_num** sustituye etiquetas verbales por una escala ordinal; **role_playing** contextualiza la tarea en un escenario empático; **cot** aplica razonamiento paso a paso (chain of thought); **cot_guiado** combina dicho razonamiento con un formato de salida estructurado; **few_shot** añade tres ejemplos ilustrativos.

Prompt reference	Accuracy	Weighted Precision	Weighted Recall	Weighted F1	Weighted Kappa
cat_simple [1]	0.5490	0.6440	0.5490	0.5253	0.5681
cat_flex_breve [2]	0.6035	0.6890	0.6035	0.6041	0.6728
cat_tol_errores [3]	0.6272	0.7044	0.6272	0.6254	0.7229
cat_definidas [4]	0.6655	0.7435	0.6655	0.6684	0.7247
escala_num [5]	0.6573	0.6741	0.6573	0.6341	0.6972
role_playing [6]	0.5610	0.6941	0.5610	0.5617	0.6197
cot [7]	0.7282	0.7691	0.7282	0.7350	0.7920
cot_guiado [8]	0.6795	0.7632	0.6795	0.6790	0.7010
few_shot [9]	0.7422	0.7533	0.7422	0.7256	0.8272

Tabla 4.5: Métricas sobre tareas de tipo 1.

Mejores desempeños: Los resultados muestran que los *prompts* **few_shot** y **cot** obtuvieron los mejores resultados en términos de *F1 ponderado* y coeficiente *Kappa ponderado*, lo cual indica un **acuerdo muy bueno** según la interpretación de esta última métrica. El *prompt* **few_shot** emplea *few-shot learning*, mientras que el **cot** sigue un enfoque de pensamiento paso a paso (*chain*

of thought). Dado el valor en las métricas mencionadas anteriormente, se infiere que estos enfoques han sido efectivos para guiar al modelo en la evaluación precisa de las respuestas.

Peores desempeños: Los *prompts* `cat_simple` y `role_playing` obtuvieron los resultados más bajos en términos de *F1 ponderado* y coeficiente Kappa ponderado. Estos *prompts* se ubicaron en la categoría de **acuerdo moderado**, lo que indica dificultades en guiar al modelo para evaluar correctamente las respuestas. El *prompt* `cat_simple` combina una categorización sin lineamientos con una breve justificación, mientras que el *prompt* `role_playing` realiza una clasificación en base a una técnica de rol que intenta ser un enfoque más empático y de apoyo en la evaluación.

Por lo tanto, al analizar las técnicas empleadas en los *prompts*, se observan ciertas tendencias significativas. Las estrategias de **pensamiento paso a paso** y **aprendizaje con ejemplos** demuestran ser efectivas para guiar al modelo y mejorar su desempeño. En contraste, enfoques basados en **categorización simple** o sin criterios claros de evaluación resultan en un rendimiento inferior.

Por otro lado, la **variabilidad** en el desempeño entre diferentes *prompts* destaca la importancia de un **diseño cuidadoso de las instrucciones**. Ya que, como se puede observar, pequeños cambios en estos pueden tener un impacto considerable en el rendimiento. Esto se evidencia al comparar *prompts* con estructuras similares. Por ejemplo, el *prompt* `cat_simple` obtuvo los peores resultados en todas las métricas, a pesar de ser muy similar a los *prompts* `cat_tol_errores` y `cat_flex_breve`. El *prompt* `cat_simple` solicita una evaluación que incluye una categoría de relevancia y pide una breve justificación. En contraste, el *prompt* `cat_tol_errores` pide la categoría de relevancia, especifica que no se incluya texto adicional y añade que se sea indulgente con repeticiones o errores menores que puedan ocurrir en respuestas habladas. El *prompt* `cat_flex_breve` también solicita relevancia y justificación, pero enfatiza que las respuestas breves pueden ser totalmente válidas.

Esto reafirma que los detalles en la formulación de los *prompts* influyen notablemente en la **calidad y consistencia** de las evaluaciones generadas.

Un caso particularmente interesante es la comparación entre los *prompts* `cat_simple` y `few_shot`. El *prompt* `few_shot` es prácticamente igual al `cat_simple` excepto por la incorporación de tres ejemplos (uno por categoría) mediante la técnica de *few-shot learning*. Esta modificación produce una mejora notable en el rendimiento, pasando de ser el peor *prompt* a uno de los mejores en términos de todas las métricas evaluadas. Es interesante notar que la simple adición de estos ejemplos logra mejores resultados que las variantes que intentan mejorar el *prompt* `cat_simple` mediante instrucciones más detalladas (como en los *prompts* `cat_tol_errores` y `cat_flex_breve`), sugiriendo que los criterios de evaluación pueden ser mejor captados a través de ejemplos que mediante descripciones explícitas.

Análisis de Matrices de Confusión: Para entender un poco más cómo se distribuyeron los errores en los mejores y peores casos se presentan, en la Figura 4.5, las matrices de confusión de los *prompts* `role_playing`, `cot`, `cat_simple` y `few_shot`. En estas matrices, las categorías 0, 1 y 2 corresponden a “Not Addressed”, “Partially Addressed” y “Fully Addressed” respectivamente.

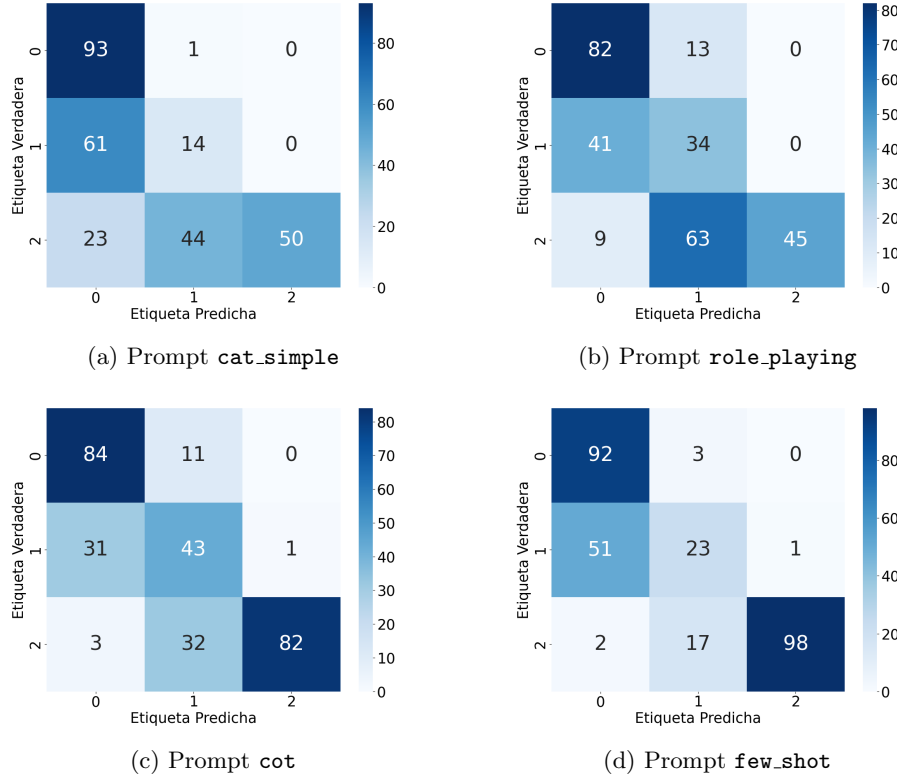


Figura 4.5: Matrices de confusión de los diferentes *prompts*.

El *prompt* `cot` muestra una buena capacidad para identificar las categorías extremas (0 y 2), con 84 y 82 predicciones correctas respectivamente. Sin embargo, presenta cierta dificultad con la categoría intermedia (1), donde hay una distribución más dispersa de predicciones.

El *prompt* `cat_simple`, que mostró el peor desempeño general, tiene una tendencia marcada a subestimar las categorías. Esto se evidencia en la cantidad significativa de casos donde el *prompt* asigna una categoría menor a la verdadera. Por ejemplo, 61 casos donde asignó categoría 0 cuando debería ser 1, y 23 casos donde asignó 0 cuando debería ser 2. Este sesgo hacia categorías más bajas podría estar relacionado con la falta de criterios claros en el *prompt*, que al solicitar simplemente una evaluación sin más guía, lleva al modelo a adoptar una postura más conservadora en sus clasificaciones.

El *prompt* **few_shot**, que incorpora few-shot learning, muestra el comportamiento más balanceado. Mantiene una alta precisión en las categorías extremas (92 y 98 predicciones correctas para 0 y 2 respectivamente) y se observa una tendencia a favorecer estas categorías por sobre la intermedia, lo que podría explicar algunas de las confusiones observadas. Se destaca como casi no comete errores entre “Not Addressed” y “Fully addressed”, con solo dos casos donde esto ocurre.

La comparación entre los *prompts* **cat_simple** y **few_shot** es particularmente interesante, ya que teniendo la misma estructura base, la inclusión de ejemplos en el *prompt* **few_shot** logra corregir significativamente el sesgo hacia categorías bajas observado en el *prompt* **cat_simple**. Esto resalta el valor de proporcionar ejemplos concretos para calibrar el criterio de evaluación del modelo como se observó anteriormente.

El *prompt* **role_playing** muestra un patrón particular donde tiende a asignar la categoría intermedia con mayor frecuencia, especialmente para respuestas que deberían ser categoría 2 (63 casos). Este comportamiento podría deberse a que el *prompt*, si bien proporciona varias instrucciones sobre el enfoque de evaluación, no establece criterios claros para distinguir entre las diferentes categorías. Ante esta ambigüedad en los criterios de clasificación, el modelo parece optar a la categoría intermedia como opción segura.

Un aspecto interesante que se observa en todos los *prompts* es que la categoría intermedia resulta la más desafiante de predecir correctamente. Esto podría explicarse por la naturaleza más subjetiva de esta categoría, donde los límites entre “parcialmente relacionado” y las categorías adyacentes pueden ser más difusos. A su vez, es destacable que no hay casos donde una respuesta haya sido evaluada como “Fully Addressed” cuando en realidad era “Not Addressed”, y apenas existen casos donde se evaluó como “Fully Addressed” siendo “Partially Addressed”. Esto sugiere que cuando el modelo califica una respuesta como “Fully Addressed”, es altamente probable que efectivamente lo sea (al menos sobre el conjunto de pruebas en el que se evaluó).

Para poder visualizar mejor el **comportamiento** de los *prompts* se procede a realizar un análisis que examina ejemplos específicos y compara las respuestas generadas por diferentes *prompts*.

Análisis de comportamiento

Este análisis se centra en observar ejemplos específicos de las evaluaciones para comprender el comportamiento y patrones de respuesta de los diferentes *prompts*.

Fortalezas de los Mejores Prompts En la mayor parte de los casos los prompts **cot** y **few_shot** mostraron un buen desempeño. A continuación se listan algunos ejemplos donde se puede observar esto.

Ante la pregunta “*What is your favourite breakfast?*” y la respuesta del estudiante “*My favorite breakfast is coffee and bread or cookies*”, el prompt **cot**

generó:

```
{
  thought_process: [
    "The student is directly answering the question by mentioning their favorite breakfast, which is a specific type of food.",
    "The answer includes multiple options, but it's still related to the question about favorite breakfast.",
    "The student's response is concise and straightforward."
  ],
  relevance: "Fully Addressed"
}
```

Se observa que el modelo sigue un razonamiento adecuado, mencionando que la respuesta está relacionada con el desayuno favorito, es directa e incluye múltiples opciones de alimentos que siguen siendo relevantes para la pregunta. La respuesta está calificada como “Fully Addressed”.

Para este caso, el prompt `few_shot` generó lo siguiente:

```
{
  relevance: "Fully Addressed",
  justification: "Direct and relevant answer to the question, listing specific breakfast items."
}
```

Observar que el modelo genera una evaluación clara y concisa sobre la respuesta del estudiante, destacando que la respuesta es directa y relevante. Justifica la decisión al observar que el estudiante lista ítems específicos de desayuno.

La elección de estos dos ejemplos está justificada por el hecho de que son casos en los que claramente la respuesta del estudiante es correcta y además profundiza un poco en su respuesta. Es importante asegurarse que los prompts utilizados clasifican bien estos casos y además admitan diferentes niveles de detalle en las respuestas.

Otro caso donde se puede ver un buen comportamiento es en la pregunta “*Can you swim?*” y la respuesta del estudiante es “*I kinda can swim. I’m not very good swimming.*”. En este caso se presenta una respuesta un poco más informal, que no sigue un formato tan simple como podría ser “*Yes, I can*” o “*No, I can’t*”, y que además deja entrever cierta duda respecto a la capacidad de nadar.

Para el prompt `cot` la respuesta fue la siguiente:

```
{
  thought_process: [
    "The question asks about the student's ability to swim, so I will look for a response that directly answers this ability.",
    "The student's answer mentions 'I kinda can swim', which implies that they have some swimming ability, but also mentions 'I'm not very good', which indicates a limitation.",
    "Considering the simplicity of the question and the brevity of
```

```

    the response, the student's answer directly addresses the
    question of their swimming ability, even if it's not a confident
    affirmation."
  ],
  relevance: "Fully Addressed"
}

```

Aquí también se puede observar que el modelo sigue un razonamiento adecuado, identificando claramente la habilidad del estudiante para nadar, aunque de manera limitada, y evaluando que la respuesta cumple con la pregunta de forma directa, a pesar de la falta de confianza en la afirmación.

El prompt `few_shot` generó lo siguiente:

```

{
  relevance: "Fully Addressed",
  justification: "Direct and relevant answer to the question, even
  though it includes additional information about the student's
  proficiency level."
}

```

De igual manera, se califica correctamente y se da una justificación adecuada en la que se destaca que la información extra acerca del nivel que presenta no compromete la clasificación de la respuesta a la pregunta.

Si bien estas técnicas muestran resultados prometedores para la evaluación del *task fulfillment*, hay casos dentro del conjunto de pruebas sobre los que se muestran inconsistencias en el criterio de respuesta. Esto provoca que evalúe bien para algunas respuestas, pero luego falle para variantes de dicha respuesta igualmente válidas. A continuación se destacan estos casos y otros en los cuales ninguno de los mejores prompts es capaz de clasificarlos correctamente.

Casos cruzados para los Mejores *Prompts* En este apartado, presentamos casos en los cuales uno de los prompts clasifica bien, pero el otro no, y viceversa.

Casos donde `cot` clasifica bien y `few_shots` no:

Para la pregunta “*What is your favourite book?*”, una posible respuesta de un estudiante es “*Las Aventuras del Sapo Ruperto*”, a la cual categorizamos como “Fully Addressed”.

La respuesta utilizando el prompt `cot` es:

```

{
  thought_process: [
    "The question asks for the student's favourite book, so I will
    look for a specific book title or name in the answer.",
    "The answer provided is a book title, but it does not provide any
    additional context or explanation, which is common in favourite
    book discussions.",
    "However, the student has provided a clear and direct response to
    the question, which is a key characteristic of a fully addressed
    answer."
  ]
}

```

```

    ],
    relevance: "Fully Addressed"
  }

```

Mientras que para `few_shot`:

```

{
  relevance: "Partially Addressed",
  justification: "Provides a specific book title but does not answer
the question about the favourite book."
}

```

En este caso, el método `cot` clasificó correctamente la respuesta “*Las Aventuras del Sapo Ruperto*” como “Fully Addressed”, reconociendo que el estudiante proporciona un título específico que responde directamente a la pregunta sobre su libro favorito. Por otro lado, `few_shot` la clasifica como “Partially Addressed”, justificando que, aunque se menciona un título, no se aclara explícitamente que sea el favorito del estudiante.

Es posible que `few_shots` esté influenciado por un sesgo hacia respuestas que incluyen una afirmación directa de preferencia o títulos ampliamente conocidos. Aunque esta razón no aparece de manera explícita en la justificación del modelo, podría contribuir a que respuestas con títulos menos reconocidos sean interpretadas como incompletas. Este caso resalta la necesidad de que los *prompts* puedan interpretar respuestas literales y específicas como completas, sin depender de que el título sea ampliamente conocido o de que la preferencia esté formulada de manera directa.

Otro caso es para la pregunta “*What is your name?*”, para la cual una respuesta de un estudiante es “*My name is Agustina, my name is Agustina, my*”, a la cual categorizamos como “Fully Addressed”.

La respuesta utilizando el *prompt cot* es:

```

{
  "thought_process": [
    "The student provided a direct answer to the question by stating 'My name is Agustina', which indicates they understand what the question is asking.",
    "The repetition of 'my name is Agustina' is not necessary and seems to be a filler or a nervous habit, but it doesn't detract from the fact that they provided the correct information.",
    "The student's response is concise and directly answers the question without any extraneous information."
  ],
  relevance: "Fully Addressed"
}

```

Mientras que para `few shots`:

```

{
  relevance: "Not Addressed",
  justification: "Answer is not relevant to the question due to repetition and unrelated content."
}

```

En este caso, el método `cot` clasificó correctamente la respuesta “*My name is Agustina, my name is Agustina, my*” como **Fully Addressed**, interpretando que, a pesar de la repetición, el estudiante proporciona una respuesta directa y completa a la pregunta sobre su nombre. La repetición es reconocida como un posible elemento de relleno que no afecta la relevancia de la respuesta.

Por otro lado, el método `few_shots` clasifica la respuesta como “Not Addressed”, justificando que la repetición y el contenido adicional hacen que la respuesta no sea relevante. Este resultado indica que en este caso el método interpreta la repetición como una falta de claridad en la respuesta, sin considerar que, en un contexto oral, esta repetición podría deberse a un esfuerzo del estudiante por asegurarse de que la información se escuche correctamente o se registre en su totalidad. Este comportamiento es común en interacciones orales, donde los estudiantes pueden repetir una frase para confirmar que su respuesta sea captada correctamente.

Además, esta clasificación como “Not Addressed” resulta problemática, ya que representa la categoría opuesta a la asignada en nuestra evaluación (“Fully Addressed”). Por ello, es importante que los *prompts* tengan en cuenta estos patrones de respuesta típicos de contextos orales, permitiendo clasificar correctamente respuestas que incluyen repeticiones sin que estas interfieran en la precisión de la información proporcionada.

Casos donde `few_shots` clasifica bien y `cot` no:

En este caso la pregunta es “*What do you like to do on weekends?*” y la respuesta del estudiante: “*Spending time with family*”. En la base de estudio, este ejemplo está categorizado como “Fully Addressed” ya que consideramos que pasar tiempo con la familia es una respuesta coherente a la pregunta.

La respuesta utilizando el prompt `cot` es:

```
{
  thought_process: [
    "The question asks about weekend activities, so I will look for a response that mentions something the student enjoys doing on weekends.",
    "The student's response is a single word, 'spending time with family', which implies a general activity rather than a specific one.",
    "The student has provided a clear and direct answer, but I need to consider whether it directly answers the question about weekend activities."
  ],
  relevance: "Partially Addressed"
}
```

En este caso, el método presenta ciertas limitaciones al interpretar la respuesta “*Spending time with family*”. Primero, identifica incorrectamente la respuesta como una única palabra (single word), cuando en realidad se trata de una frase que responde plenamente a la consigna. Además, el método clasifica la actividad como general, lo cual ignora el contexto de la pregunta.

Para este mismo ejemplo, el prompt `few_shot` generó un resultado correcto, siendo la salida:

```
{
  relevance: "Fully Addressed",
  justification: "Directly answers the question about weekend
activities."
}
```

Considera que la respuesta responde directamente a la pregunta.

Otro caso en el que `few_shots` da bien y `cot` no, es el caso de la pregunta “*What is your favourite breakfast?*” y la respuesta del estudiante “*Coffe with milk*”, a la cual clasificamos como “Fully Addressed”.

La respuesta utilizando el prompt `cot` es:

```
{
  thought_process: [
    "The student has mentioned a specific food item, which is 'coffee
with milk', but it is typically consumed at any time of day, not
necessarily for breakfast.",
    "The question asks for the student's favourite breakfast,
implying a meal that is eaten in the morning.",
    "Although coffee with milk is not a traditional breakfast food,
it is still a possible breakfast choice, even if it's not the
most common one."
  ],
  relevance: "Partially Addressed"
}
```

En este caso, el método clasifica la respuesta como “Partially Addressed” posiblemente debido a un sesgo cultural implícito. La justificación en el proceso de pensamiento sugiere que el modelo no considera “café con leche” como una opción válida de desayuno, al interpretarlo como una bebida que se puede consumir en cualquier momento del día y no necesariamente en el desayuno. Sin embargo, en el contexto cultural de Uruguay, “café con leche” es una opción habitual de desayuno y cumple plenamente con la consigna dada. Este ejemplo resalta la importancia de adaptar los criterios de evaluación para, de alguna forma, reflejar las prácticas y preferencias locales, evitando así clasificaciones erróneas derivadas de sesgos culturales en el modelo, como en este caso.

Para este ejemplo, el prompt con técnica `few_shots` generó el resultado:

```
{
  relevance: "Fully Addressed",
  justification: "Direct and relevant answer to the question about
breakfast preferences."
}
```

Esta respuesta es correcta y directa.

Existe otra versión de respuesta de estudiante a esta misma pregunta, en la cual en vez de decir “*coffe with milk*” o “*My favorite breakfast is coffee and bread or cookies*”, dice solo “*coffe*”. En este caso, ambos prompts lo clasificaron como “Partially Addressed”, aunque la etiqueta designada es “Fully Addressed”. Para

este caso, la justificación de los modelos es similar, argumentando que café solo no es una comida válida para desayuno.

En el siguiente apartado veremos otros casos en los que fallan ambos de nuestros mejores prompts hasta el momento.

Casos Negativos para los Mejores *Prompts* En el caso de la pregunta “*Do you prefer cats or dogs*” y la respuesta del estudiante “*My favourite animal is a dog*”, a la cual clasificamos como “Fully Addressed”, ambos prompts la clasificaron como “Partially Addressed”.

El prompt con técnica *cot*:

```
{
  thought_process: [
    "The question asks for a preference between cats and dogs, so I'll look for a direct answer to this specific question.",
    "The student mentions their favourite animal is a dog, but doesn't explicitly state a preference between cats and dogs.",
    "The student's answer implies that they have a strong affinity for dogs, but doesn't provide a direct comparison with cats."
  ],
  relevance: "Partially Addressed"
}
```

Mientras que el prompt *few_shot*:

```
{
  relevance: "Partially Addressed",
  justification: "The student answers the question indirectly by mentioning their favourite animal, but does not explicitly state a preference between cats or dogs."
}
```

En este caso, ambos métodos clasifican la respuesta “*My favourite animal is a dog*” como “Partially Addressed” al interpretar que el estudiante no expresa explícitamente una preferencia entre gatos y perros, sino que menciona un animal favorito. Sin embargo, se considera que esta afirmación refleja una preferencia implícita por los perros, dado que el estudiante los destaca al responder a una pregunta sobre preferencias entre ambas mascotas. Este tipo de respuesta es válida en contextos informales y cumple con la consigna sin necesidad de una comparación explícita. La discrepancia sugiere que es necesario un mayor refinamiento en los prompts, ya que deberían reconocer este tipo de respuestas implícitas como completas, permitiendo así una interpretación más flexible en preguntas donde no se requiere una afirmación directa.

En contraste, con la respuesta “*I prefer dogs*”, ambos métodos la clasificaron correctamente como “Fully Addressed”, lo cual indica que son capaces de clasificar respuestas directas y explícitas a esta pregunta.

El caso a continuación, para la pregunta “*What time do you go to school?*” y la respuesta del estudiante “*I go to school at 7pm*”, a la cual clasificamos como “Fully Addressed”, ambos prompts la clasificaron como “Partially Addressed”.

Para el prompt cot:

```
{
  thought_process: [
    "The student is directly answering the question about the time they go to school.",
    "However, the answer is unusual as school hours are typically during the day, not in the evening.",
    "Considering the question is about the time, the answer still provides a specific time, even if it's not typical.",
    "I am looking for a relevant answer based on the context of the question, which is about the time of day they go to school."
  ],
  relevance: "Partially Addressed"
}
```

El prompt few shots:

```
{
  relevance: "Partially Addressed",
  justification: "Provides a specific time but it's not a typical time for school, as most schools operate during the day."
}
```

En este ejemplo, ambos métodos clasificaron la respuesta “I go to school at 7pm” como “Partially Addressed”, dado que el horario mencionado es inusual, según lo indicado explícitamente en su proceso de pensamiento. Ambos prompts asumen que el horario escolar suele ser durante el día, por lo que interpretan la respuesta como incompleta. Sin embargo, la respuesta del estudiante proporciona un horario específico, lo que cumple directamente con la consigna de la pregunta, independientemente de si el horario es típico o no. Este caso resalta la importancia de aplicar un criterio que permita a los prompts considerar como válidas las respuestas en función de la precisión de la información proporcionada, más allá de si coincide con expectativas convencionales. Así, los prompts podrían reconocer respuestas completas en situaciones donde el estudiante responde literalmente a la pregunta, sin ajustar su respuesta a supuestos contextuales.

Combinación de *Few-Shot Learning* y *Chain of Thought* Dado que las técnicas de *Few-Shot Learning* y *Chain of Thought* han demostrado mejorar individualmente el desempeño del modelo, en este apartado proponemos combinarlas para intentar aprovechar sus fortalezas. El *prompt* se encuentra en el Anexo [10].

Al utilizar este *prompt* combinado para hacer inferencia sobre nuestra base de datos, se obtiene un valor de 0,8308 para *Weighted Kappa* y 0,7522 en *Weighted F1*. Esto es una mejora sutil respecto a los valores obtenidos en estas métricas para las técnicas por separado.

En la Figura 4.6 se presenta la matriz de confusión para este caso. Al compararla con las presentadas en la Figura 4.5 se observa que la combinación de técnicas genera una tendencia hacia clasificaciones más altas. Si bien esto introduce algunos casos donde respuestas son incorrectamente clasificadas como

“Fully Addressed”, también fortalece la diagonal principal de la matriz, logrando una mejor identificación de las respuestas que efectivamente abordan completamente la tarea.

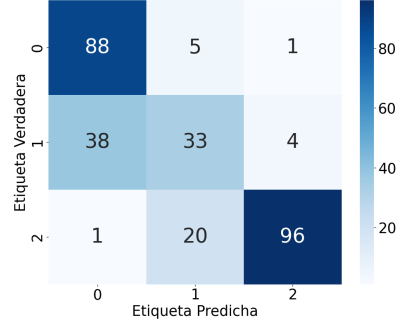


Figura 4.6: *Prompt chain of thought + a few shots* para tareas tipo 1.

4.8.2. Análisis para Tareas de Tipo 2

Se presenta el análisis de los resultados obtenidos para las tareas de tipo 2, abordando tanto aspectos cuantitativos como cualitativos. Los *prompts* específicos utilizados se encuentran detallados en el Anexo A.4.2. Recordar que para este tipo de tareas, se trata de evaluar respuestas a preguntas simples sobre imágenes.

Análisis sobre métricas obtenidas

En la Tabla 4.6 se pueden observar los resultados obtenidos por cada tipo de *prompt*, reportado en las distintas métricas definidas.

Prompt reference	Accuracy	Weighted Precision	Weighted Recall	Weighted F1	Weighted Kappa
cot_libre [1]	0.5938	0.6439	0.5938	0.6143	0.6174
cot_estructurado [2]	0.7083	0.8005	0.7083	0.7326	0.7614
few_shot_3 [3]	0.7604	0.7644	0.7604	0.7601	0.8158
few_shot_7 [4]	0.7708	0.7750	0.7708	0.7721	0.8025
cot_fewshot_3 [5]	0.7955	0.7761	0.7955	0.7721	0.8030
contrastive [6]	0.6979	0.7296	0.6979	0.7104	0.7087

Tabla 4.6: Métricas sobre tareas de tipo 2

Mejores desempeños: Los resultados indican que los *prompts* *few_shot_3*, *cot_fewshot_3* y *few_shot_7* obtuvieron los mejores resultados, todos con un coeficiente *Kappa ponderado* superior a 0.8 (0.8158, 0.8030 y 0.8025 respectivamente), lo cual indica un **muy buen acuerdo** según la interpretación de esta métrica. Es destacable que los tres mejores *prompts* incorporan la técnica de

few-shot learning, variando en la cantidad de ejemplos (3 o 7) y en la combinación con otras técnicas como *chain of thought*. El *prompt few_shot_3* alcanza el mejor rendimiento general con un *kappa* de 0.8158, seguido muy de cerca por los otros dos enfoques. El alto rendimiento de estos tres *prompts* refuerza que la incorporación de ejemplos es particularmente efectiva para guiar al modelo en la evaluación de respuestas basadas en imágenes, independientemente de si se combinan con otras técnicas o se varía la cantidad de ejemplos utilizados. Notar que, a pesar de probar un enfoque con 7 ejemplos, este no mejora al de 3 ejemplos en los resultados obtenidos.

Peores desempeños: El *prompt cot_libre* obtuvo los resultados más bajos en todas las métricas evaluadas, con un *F1 ponderado* de 0.6143 y un coeficiente *kappa ponderado* de 0.6174. Este *prompt* utiliza un enfoque de *chain of thought* libre, sin ejemplos ni guía estructurada, lo que parece ser insuficiente para la complejidad inherente a la evaluación comparativa contra un contexto de referencia.

Análisis de Matrices de Confusión: Para profundizar en la comprensión de los patrones de error, se analizan las matrices de confusión mostradas en la Figura 4.7. En estas matrices, las categorías 0, 1 y 2 corresponden a “Not Addressed”, “Partially Addressed” y “Fully Addressed” respectivamente.

El *prompt few_shot_3* muestra buenos resultados en su distribución de predicciones. Logra identificar correctamente 35 respuestas en la categoría 0 y 31 en la categoría 2. Es importante notar que casi no comete errores graves, con solo 2 casos donde clasificó como correcta una respuesta que no abordaba el tema. En la categoría intermedia obtiene 7 aciertos, aunque muestra algunas confusiones al clasificar 6 casos como no relacionados y 4 como totalmente correctos.

El *prompt cot_libre* muestra dificultades en su desempeño. Si bien identifica correctamente 27 casos en categoría 0, tiene una tendencia marcada a confundir categorías adyacentes. Esto se ve especialmente en la categoría intermedia, donde solo acierta 5 casos, y en la categoría 2, donde solo identifica correctamente 25 respuestas. Además, comete varios errores graves, con 5 casos donde clasificó como totalmente correctas respuestas que no abordaban el tema, y otros 5 casos donde las clasificó como parcialmente correctas.

El *prompt cot_fewshot_3*, que combina pensamiento paso a paso con ejemplos, muestra buenos resultados en su distribución de predicciones. Logra identificar correctamente 28 casos en categoría 0 y 38 en categoría 2. Se destaca que casi no comete errores graves, sin casos donde clasifique como totalmente incorrecta una respuesta correcta, y solo 4 casos donde clasificó como totalmente correcta una respuesta que no abordaba el tema. En la categoría intermedia obtiene 4 aciertos, con una distribución balanceada de errores.

Todos los *prompts* muestran dificultades para manejar la categoría interme-

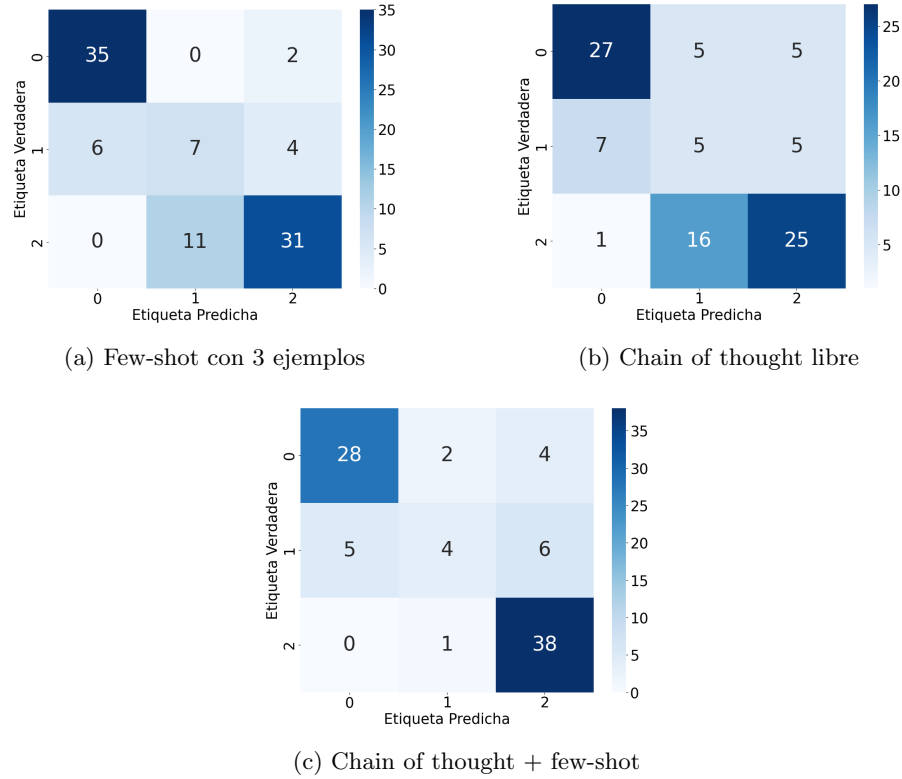


Figura 4.7: Matrices de confusión para tareas tipo 2.

dia “Partially Addressed”. Esto se refleja en predicciones más dispersas y menor precisión que en las categorías extremas. Al igual que en las tareas de tipo 1, esto sugiere que evaluar respuestas parcialmente correctas es un desafío común por la dificultad de transmitir los matices de evaluación. De todas formas para poder concluir con certeza esto, sería necesario evaluar más cantidad de casos etiquetados como “Partially Addressed”.

Análisis sobre métricas obtenidas

Para las técnicas *chain of thought* se estudiará el caso de dos prompts, *cot_libre* y *cot_estructurado*, obteniendo para estos casos como máximo en la métrica *Weighted Kappa* un valor de 0.76.

Por ejemplo para la pregunta: “*What are they?*”, que corresponde a una imagen de unos lentes de sol, con respuesta de estudiante “*they are a pair of sunglasses*” y cuya etiqueta en nuestra base es “Fully Addressed”, el *prompt cot_libre* lo clasifica incorrectamente como “Partially Addressed”.

```

{
  thought_process: [
    "The student's answer is very brief and doesn't provide any details about the sunglasses.",
    "However, it does acknowledge the main object in the image, which is a pair of sunglasses.",
    "Considering the simplicity of the answer, it's likely that the student focused on the core part of the question."
  ],
  extra_content: {
    present: true,
    details: [
      "The description of the sunglasses as 'classic', 'sleek', 'black metal frame', 'large and tinted dark grey lenses', 'thin, curved temple arms', and 'double bridge' are not mentioned by the student."
    ]
  },
  relevance: "Partially Addressed"
}

```

Si bien se puede notar que al comienzo del proceso de pensamiento el modelo identifica que la respuesta contiene información sobre el objeto principal en la imagen y se enfoca en eso, luego justifica que en la respuesta no se menciona ningún detalle adicional y luego lo clasifica como “Partially Addressed”.

Esto es un error grave ya que se infiere que generalmente las respuestas no contendrán tanto detalle adicional y es necesario asegurarse de que sea posible clasificarlas como correctas de igual manera.

En la variante `cot_estructurado`, este ejemplo se clasifica correctamente. La diferencia entre ambos prompts se centra en cómo cada uno aborda los detalles adicionales. El primer prompt (`cot_libre`) sigue un análisis con pasos no tan definidos, evaluando la presencia del objeto principal en la respuesta, a la vez que la ausencia de detalles específicos. Esta falta de definición lo lleva a penalizar respuestas breves. En contraste, el segundo prompt (`cot_estructurado`) define más claramente los pasos, lo cual le permite seguir de mejor manera los lineamientos dados.

Paralelamente, se implementaron técnicas few shots. Particularmente se aborda el análisis de dos casos: uno en el cual se construye el prompt con 3 ejemplos (`few_shot_3`) y el otro 7 ejemplos (`few_shot_7`)

Para el prompt de 3 ejemplos, se tiene que clasifica correctamente el par pregunta-respuesta presentado anteriormente, de los lentes de sol. Además, es el único prompt que clasifica correctamente el par pregunta-respuesta “*What animal is this? it is a brown horse with a long tail*”, correspondiente a una imagen con la descripción: *A dark brown horse with a black mane and tail. It is galloping in a field with grass, and there is a fence and trees in the blurred background.*

```
{
  content: {
    explanation: "Student correctly identified the main animal and a
    key characteristic",
    extra_content: {
      present: true,
      details: "unverifiable detail about tail length"
    },
    relevance: "Fully Addressed"
  }
}
```

Esta respuesta logra identificar que el largo de la cola del caballo no está presente en la descripción, pero no penaliza por eso.

El prompt `few_shot_7` logra un comportamiento similar en la mayoría de los casos evaluados, incluso empeorando en algunos, lo cual deja en evidencia que más ejemplos no necesariamente implican un mejor desempeño. Ejemplos como el mencionado anteriormente para el prompt `few_shot_3` pasaron a ser erróneamente clasificados, quizás producto de la inclusión de ejemplos que tienden a confundir al modelo en su evaluación. En cambio, el prompt `few_shot_3` es más directo, proporcionando pocos ejemplos, pero significativos, los cuales le ayudan al modelo a seguir un criterio adecuado de evaluación. Sin embargo, se debe ser cauteloso al interpretar estos resultados. El tamaño limitado de la base de datos y su relativamente baja variabilidad no nos permiten descartar por completo los beneficios de incluir más ejemplos en el prompt. Por ejemplo, si se tiene un ejemplo que representa un caso específico, pero poco frecuente en nuestra base de datos, su inclusión podría no mostrar mejoras significativas en las métricas, simplemente porque no hay suficientes casos similares para evaluar su impacto real.

Luego, análogamente al estudio del tipo de tarea anterior nos propusimos generar un prompt aplicando ambas técnicas, few shots y chain of thought (`cot_fewshot_3`). En la prueba realizada, la métrica Weighted Kappa para este caso es 0.803.

Este prompt es el único que logra clasificar correctamente una pregunta que parte de la imagen de un helado, dice: “What food is it?” y la respuesta del estudiante es “Ice cream with chocolate sprinkles on top”. En particular, la dificultad de este ejemplo es que la información de “with chocolate sprinkles on top” no está contenida en la descripción de la imagen, pero tampoco nada que lo contradiga. Es por ello que debería ser clasificada con “Fully Addressed” y este prompt lo logra.

Adicionalmente, se realizó otro prompt, al que llamamos enfoque contrastivo (prompt `contrastive`) debido a que sigue la siguiente indicación:

```
{
  instructions: "Evaluate the student's answer by contrasting it with
  two extreme scenarios:\n1. A perfect answer that fully addresses the
  question based on the image description.\n2. An entirely unrelated
```

```

    answer that doesn't address the question at all.",
    contrastive_analysis: {
        similarity_to_perfect: score,
        similarity_to_unrelated: score
    },
    relevance: category,
    explanation: "brief justification"
}

Score similarities from 0 (not similar at all) to 10 (very similar).
Use one of these categories: "Not Addressed", "Partially Addressed", or "
Fully Addressed".

```

Para el par pregunta-respuesta “*What is she doing?*” “*Working on a computer*”, y una imagen con descripción: “*A female doctor sitting at a desk, working on a laptop. She is wearing a white lab coat with a stethoscope around her neck and glasses. The doctor appears focused on her work. In the background, there is a clock and a poster with medical diagrams. The office has a clean and professional appearance*”.

La salida del modelo resulta:

```

{
  contrastive_analysis: {
    similarity_to_perfect: 8,
    similarity_to_unrelated: 0
  },
  relevance: "Fully Addressed",
  explanation: "The student's answer directly addresses the question by describing the primary action depicted in the image, which is the doctor working on a laptop. The answer is brief and to the point, and while it doesn't mention the context or details, it accurately conveys the main action. The similarity to a perfect answer is high because the student has identified the key action in the image, but the score is not 10 because the answer lacks additional details. The similarity to an unrelated answer is low because the student's response is directly tied to the image description."
}

```

Para este prompt, la métrica Weighted Kappa da 0,709.

En conclusión, los mejores resultados en las tareas de tipo 2 se lograron utilizando *prompts* que incorporan *few shot learning*, ya sea solos o combinados con *chain of thought*. Los *prompts* `few_shot_3`, `cot_fewshot_3` y `few_shot_7` obtuvieron coeficientes *Kappa ponderados* superiores a 0.8, evidenciando un acuerdo excelente. Esto demuestra la eficacia de incluir ejemplos específicos para guiar al modelo en la evaluación de respuestas basadas en imágenes.

Sin embargo, todos los *prompts* mostraron dificultades para manejar la categoría intermedia “Partially Addressed” lo que indica un desafío en la evaluación automatizada de respuestas parcialmente correctas. Consideramos que un refinamiento adicional de los *prompts*, enfocándose en ejemplos más claros y en la optimización de su diseño, podría mejorar aún más los resultados y reducir errores en las categorías intermedias.

4.8.3. Análisis para Tareas de Tipo 3

A continuación, se presenta el análisis de los resultados obtenidos utilizando el enfoque de *chain of thought* con análisis estructurado. Los *prompts* específicos utilizados se encuentran detallados en el Anexo A.4.3.

Análisis sobre métricas obtenidas

Las metricas obtenidas, presentadas en la Tabla 4.7, muestran un gran desempeño del enfoque `chain_of_thought`, alcanzando un coeficiente *kappa ponderado* de 0,8607. Este valor indica un acuerdo muy bueno entre las evaluaciones del modelo y las etiquetas de referencia. El *F1 ponderado* de 0,8206 refleja un buen balance considerando la complejidad añadida de evaluar múltiples criterios simultáneamente en las respuestas.

Prompt reference	Accuracy	Weighted Precision	Weighted Recall	Weighted F1	Weighted Kappa
chain_of_thought [1]	0.82	0.8345	0.82	0.8206	0.8607
multi-part [2]	0.7170	0.7378	0.7170	0.6664	0.7549

Tabla 4.7: Métricas sobre tareas de tipo 3.

Por otra parte, el enfoque basado en una evaluación por partes (`multi-part`), mostró un acuerdo bueno. Es notable que la diferencia en rendimiento entre ambos enfoques se acentúa particularmente en el *F1 ponderado*, donde se observa la mayor diferencia. Esta métrica, al ser más sensible a los desbalances entre precisión y recall, sugiere que el enfoque `chain_of_thought` es especialmente efectivo en mantener un rendimiento consistente a través de todas las clases de clasificación, mientras que el enfoque `multi-part` muestra mayor variabilidad en su capacidad predictiva.

Es importante notar que, contrario a lo que podría esperarse, el enfoque secuencial guiado (*multi-part*) no superó al enfoque integrado de `chain_of_thought`, a pesar de que ambos buscan estructurar explícitamente el análisis de criterios. Los resultados muestran una diferencia significativa en rendimiento (F1 ponderado de 0.8206 versus 0.6664), sugiriendo que dividir el razonamiento en etapas discretas no necesariamente facilita una evaluación más precisa. Sin embargo, es necesario contextualizar estos resultados considerando las limitaciones del conjunto de evaluación: se analizaron 55 respuestas derivadas de 9 ejemplos de tareas base, donde se variaron los criterios requeridos para generar diferentes instancias de evaluación. Este tamaño de muestra, si bien permite observar diferencias claras en el rendimiento, sugiere la necesidad de validación adicional con un conjunto más amplio de casos.

Para comprender mejor estos resultados cuantitativos, en la siguiente sección se presenta un análisis que examina casos específicos, ilustrando cómo el mejor enfoque maneja diferentes aspectos de la evaluación. Nos centraremos solo en el prompt `chain_of_thought`.

Análisis de comportamiento

Casos Exitosos El modelo muestra particular fortaleza en la evaluación de respuestas detalladas que abordan múltiples aspectos de la tarea. Por ejemplo, para la pregunta *“Tell me about a party you enjoyed”* con la respuesta:

“I loved my birthday party at home. My friends came and we ate cake. I really liked the games we played.”

El modelo genera el siguiente análisis:

```
{
  reasoning: {
    step1_core: {
      on_topic: true,
      meaningful: true,
      analysis: "The response is about a party the student enjoyed,
        addressing the main task directly."
    },
    step2_criteria: {
      required: ["The place and the people", "What you enjoyed most about
        it", "What you were celebrating"],
      met_explicit: ["The place (home)", "What you enjoyed most about it
        (games)", "What you were celebrating (birthday)"],
      met_implicit: ["The people (friends came)"],
      truly_missing: []
    },
    step3_coherence: {
      logical: true,
      contradictions: "none"
    },
    step4_decision: "The response is Fully Addressed as it meets all the
      required criteria and provides a meaningful answer."
  },
  assessment: {
    explanation: "The student's response directly addresses the main task
      and meets all the required criteria, including explicitly stating
      the place, people, and what they enjoyed, as well as implicitly
      mentioning the people (friends) and what they were celebrating (
      birthday).",
    relevance: "Fully Addressed"
  }
}
```

Este caso muestra la capacidad del modelo para identificar criterios cumplidos tanto explícita como implícitamente y evaluar la coherencia global de la respuesta, reconociendo múltiples elementos que contribuyen a cumplir un mismo criterio.

Otro ejemplo exitoso se observa en la evaluación de la pregunta *“Tell me about your favourite famous person”* con la respuesta:

“Messi is a football player. He’s short and very fast.”

El modelo analiza:

```

{
  reasoning: {
    step1_core: {
      on_topic: true,
      meaningful: true,
      analysis: "The response provides a meaningful answer about
        Messi's occupation and characteristics."
    },
    step2_criteria: {
      required: ["Occupation", "Physical description"],
      met_explicit: ["Occupation"],
      met_implicit: ["Physical description"],
      truly_missing: []
    },
    step3_coherence: {
      logical: true,
      contradictions: "none"
    },
    step4_decision: "The response is fully addressed as it meets all
      required criteria, with both occupation and physical
      characteristics being provided."
  }
}

```

Este caso muestra la capacidad del modelo para evaluar respuestas más concisas pero igualmente efectivas, reconociendo que la brevedad no implica necesariamente una respuesta incompleta.

Casos Problemáticos Sin embargo, el modelo también muestra algunas limitaciones y casos donde su evaluación podría mejorar. Por ejemplo, para la pregunta “Tell me about a party you enjoyed”, en la respuesta:

“The New Year’s party was at the beach. We danced and watched fireworks. I was there with my family”

El modelo la clasifica como “Partially Addressed” ya que el paso 2 dice lo siguiente:

```

{
  reasoning: {
    step2_criteria: {
      required: ["The place and the people", "What you enjoyed most
        about it", "What you were celebrating"],
      met_explicit: ["The place (beach)", "The people (family)"],
      met_implicit: ["What you enjoyed most about it (dancing and
        watching fireworks)"],
      truly_missing: ["What you were celebrating (New Year's party,
        but not explicitly stated that it was a celebration)"]
    }
  }
}

```

Este caso revela una limitación del modelo al ser demasiado estricto con la interpretación de criterios implícitos. A pesar de que “New Year’s party” claramente implica qué se estaba celebrando, el modelo requiere una mención más

explícita.

Otro caso problemático aparece con respuestas mínimas como:

“I went to a party. There was pretty nice music.”

El modelo correctamente la clasifica como “Not Addressed”, pero su razonamiento podría ser más preciso en cuanto a por qué la respuesta es insuficiente:

```
{
  reasoning: {
    step1_core: {
      on_topic: true,
      meaningful: false,
      analysis: "The response mentions a party, but does not
        provide any specific details or meaningful information about
        it."
    },
    step2_criteria: {
      truly_missing: ["The place and the people", "What you enjoyed
        most about it", "What you were celebrating"]
    }
  }
}
```

Limitaciones Identificadas El análisis cualitativo revela algunas limitaciones consistentes en el modelo. Primero, muestra rigidez al evaluar criterios implícitos; puede ser demasiado estricto y, en ocasiones, no reconoce que ciertos términos o frases implican el cumplimiento de un criterio. Por ejemplo, “New Year’s party” implica “What you were celebrating”, pero el modelo puede no captar esta implicación.

Además, se observan inconsistencias en la clasificación de respuestas que cumplen con los criterios de manera similar, lo que sugiere una falta de consistencia en la aplicación de las reglas. Esta variabilidad en las evaluaciones indica que es necesario mejorar la uniformidad del modelo.

Finalmente, el modelo podría mejorar en la evaluación de respuestas que proporcionan información parcial pero relevante. El manejo adecuado de información parcial es esencial para ofrecer evaluaciones más precisas y completas.

Los resultados sugieren que el enfoque de *chain of thought* proporciona un marco robusto para la evaluación de tareas tipo 3, particularmente debido a la evaluación sistemática y transparente de cada criterio, la capacidad de distinguir entre cumplimiento explícito e implícito, y el análisis estructurado que facilita la comprensión del razonamiento. Sin embargo, se identifican áreas de mejora: el modelo podría beneficiarse de una mayor flexibilidad para reconocer implicaciones contextuales evidentes, es necesario mejorar la consistencia en la aplicación de las reglas de evaluación para reducir variaciones en respuestas similares, y el modelo debería ajustar su evaluación para reconocer que respuestas breves pueden ser completas si cumplen con los criterios requeridos.

4.8.4. Análisis para Tareas de Tipo 4

En esta sección se presenta el análisis de los resultados obtenidos en las tareas de tipo 4. Para esta evaluación se diseñó un único *prompt*, detallado en el Anexo A.4.4. Dado el conjunto limitado de respuestas disponibles sobre una única tarea de tarea, optamos por un análisis cualitativo, enfocándonos en ejemplos específicos que ilustran el comportamiento del sistema ante la tarea propuesta. Destacar cómo el modelo recibe como entrada la lista de diferencias válidas y dos valores numéricos, la cantidad máxima de errores admitidos y la cantidad mínima de diferencias requeridas. En particular para la tarea en la que se presenta la evaluación se tomó 2 como la cantidad máxima de errores y 2 como la cantidad mínima de aciertos para tener “Fully Addressed”, pero al menos haya una bien debería dar “Partially Addressed” y en caso de no mencionar ninguna bien sería “Not Addressed”.

Análisis de comportamiento

Este análisis se basa en examinar casos específicos donde el modelo mostró fortalezas y debilidades al evaluar las respuestas de los estudiantes. Los ejemplos incluyen tanto casos positivos, donde el modelo actuó correctamente, como casos negativos, donde se observaron errores o inconsistencias.

Fortalezas del Modelo En varios casos, el modelo demostró una capacidad adecuada para evaluar las respuestas de los estudiantes según las pautas establecidas. A continuación, se presentan algunos ejemplos donde el modelo realizó una evaluación correcta y justificó adecuadamente sus decisiones.

Por ejemplo, ante la respuesta del estudiante:

“In A have one tennis ball and B got two tennis ball. In A we can see one window with forest, but B have two window with mountains. A is day with sun light and B is night with moon. Teacher in A no have hair and look angry, but in B he got beard and look more happy.”

```
{
  analysis: {
    valid_differences_found: [
      {
        difference: "Tennis balls",
        mentioned: true,
        evidence: "In A have one tennis ball and B got two tennis ball.",
        notes: ""
      },
      {
        difference: "Time of day",
        mentioned: true,
        evidence: "A is day with sun light and B is night with moon.",
        notes: ""
      },
    ],
  },
}
```

```

    {
      difference: "Windows",
      mentioned: true,
      evidence: "In A we can see one window with forest, but B
      have two window with mountains.",
      notes: ""
    },
    {
      difference: "Window view",
      mentioned: true,
      evidence: "In A we can see one window with forest, but B
      have two window with mountains.",
      notes: ""
    },
    {
      difference: "Sports teacher",
      mentioned: true,
      evidence: "Teacher in A no have hair and look angry, but
      in B he got beard and look more happy.",
      notes: ""
    }
  ],
  valid_differences_not_found: [
    {
      difference: "Student activities"
    },
    {
      difference: "Trophy display"
    },
    {
      difference: "Student distribution"
    }
  ],
  incorrect_statements: [],
  counts: {
    correct_differences: 5,
    actual_errors: 0
  },
  requirements_check: {
    meets_minimum: true,
    within_error_limit: true
  }
},
assessment: {
  category: "Fully Addressed",
  explanation: "The student has met the minimum required differences
  and has not exceeded the maximum allowed errors."
}
}

```

Aquí, el modelo reconoció correctamente las diferencias mencionadas, a pesar de las faltas gramaticales y el uso de lenguaje simple, tal como se indica en las pautas de evaluación. Esto muestra la capacidad del modelo para interpretar respuestas de estudiantes de inglés como segundo idioma y evaluarlas adecuadamente.

Otro caso: *“The teacher has a different appearance in both images. The lighting conditions are different between day and night. There appears to be a different arrangement of students in the space.”*

```
{
  analysis: {
    valid_differences_found: [
      {
        difference: "Sports teacher",
        mentioned: true,
        evidence: "The teacher has a different appearance in both images.",
        notes: "Student correctly identified a difference in the teacher's appearance."
      },
      {
        difference: "Time of day",
        mentioned: true,
        evidence: "The lighting conditions are different between day and night.",
        notes: "Student correctly identified a difference in the time of day."
      },
      {
        difference: "Student distribution",
        mentioned: true,
        evidence: "There appears to be a different arrangement of students in the space.",
        notes: "Student correctly identified a difference in student distribution."
      }
    ],
    valid_differences_not_found: [
      {
        difference: "Tennis balls"
      },
      {
        difference: "Windows"
      },
      {
        difference: "Window view"
      },
      {
        difference: "Trophy display"
      },
      {
        difference: "Student activities"
      }
    ],
    incorrect_statements: [],
    counts: {
      correct_differences: 3,
      actual_errors: 0
    },
    requirements_check: {
      meets_minimum: true,
      within_error_limit: true
    }
  }
}
```

```

    }
  },
  assessment: {
    category: "Fully Addressed",
    explanation: "Student identified 3 valid differences, met the
    minimum required differences, and had no actual errors."
  }
}

```

Notar que en este caso el modelo también reconoció correctamente. Esta respuesta, si bien no contiene errores gramaticales, es menos estructurada que la otra. Aún así logra interpretar y extraer con éxito la respuesta correcta.

Casos donde el Modelo Presentó Dificultades Aunque el modelo mostró fortalezas, también hubo casos donde su desempeño no fue el esperado. A continuación, analizamos algunos ejemplos donde el modelo no clasificó correctamente las respuestas o cometió errores en su análisis.

Por ejemplo, ante la respuesta:

“A got one tennis ball and B got two. Window showing different things A got forest and B got mountains. Floor is different color. Teacher look different because A no hair and B have beard.”

```

{
  analysis: {
    valid_differences_found: [
      {
        difference: "Tennis balls",
        mentioned: true,
        evidence: "A got one tennis ball and B got two.",
        notes: "Simple language used"
      },
      {
        difference: "Window view",
        mentioned: true,
        evidence: "Window showing different things - A got forest
        and B got mountains.",
        notes: "Simple language used"
      },
      {
        difference: "Sports teacher",
        mentioned: true,
        evidence: "Teacher look different because A no hair and B
        have beard.",
        notes: "Simple language used, minor grammatical error"
      }
    ],
    valid_differences_not_found: [
      {
        difference: "Time of day"
      },
      {
        difference: "Windows"
      },
      {
        difference: "Student activities"
      }
    ]
  }
}

```

```

    },
    {
      difference: "Trophy display"
    },
    {
      difference: "Student distribution"
    }
  ],
  incorrect_statements: [
    {
      stated: "Floor is different color.",
      why_incorrect: "This difference does not exist in the
        reference list and contradicts the actual images."
    }
  ],
  counts: {
    correct_differences: 3,
    actual_errors: 1
  },
  requirements_check: {
    meets_minimum: true,
    within_error_limit: true
  }
},
assessment: {
  category: "Partially Addressed",
  explanation: "The student identified three valid differences but
    missed five and made one error. Although they meet the minimum
    required differences, they exceed the maximum allowed errors."
}
}

```

En este caso, aunque el estudiante indicó suficientes diferencias válidas y solo cometió un error (la mención del color del piso), el modelo clasificó la respuesta como “Partially Addressed” en lugar de “Fully Addressed”. Según las pautas, el máximo de errores permitidos es 2, y el estudiante solo cometió uno, por lo que debería haberse clasificado como “Fully Addressed”. Esto indica una inconsistencia en la aplicación de los criterios de evaluación por parte del modelo. Notar que de igual manera sí detectó correctamente todos los aciertos y errores, clasificándolos adecuadamente en las listas correspondientes, incluso argumentando de igual manera.

El fallo ocurrió en el último paso al dar la calificación, lo cual puede deberse a una interpretación errónea de sus propios resultados, ya que incluso teniendo los “requirements_check” con valores “meets_minimum: true” y “within_error_limit: true”, el modelo ignoró estas validaciones exitosas al momento de asignar la categoría final. Este caso se repitió con otras respuestas. Dado que el último chequeo es más que nada una validación lógica, si se diera que la separación en diferencias válidas e inválidas las hace bien, podría usarse esta salida como respuesta, y calcular por fuera el *task fulfillment* dependiendo de la cantidad de respuestas en cada lista.

Otro ejemplo donde el modelo mostró dificultades es la respuesta:

“Thing in room is different place. Everything look not same. Activity is changed. All stuff is moved other place.”

```
{
  analysis: {
    valid_differences_found: [
      {
        difference: "Student activities",
        mentioned: true,
        evidence: "Activity is changed",
        notes: "The student mentions that the activity has changed ."
      },
      {
        difference: "Student distribution",
        mentioned: true,
        evidence: "All stuff is moved other place",
        notes: "The student implies that the students are in different locations."
      }
    ],
    valid_differences_not_found: [
      {
        difference: "Tennis balls"
      },
      {
        difference: "Time of day"
      },
      {
        difference: "Windows"
      },
      {
        difference: "Window view"
      },
      {
        difference: "Sports teacher"
      },
      {
        difference: "Trophy display"
      }
    ],
    incorrect_statements: [
      {
        stated: "Thing in room is different place. Everything look not same.",
        why_incorrect: "These statements are too vague and do not correspond to specific differences."
      }
    ],
    counts: {
      correct_differences: 2,
      actual_errors: 1
    },
    requirements_check: {
      meets_minimum: true,
      within_error_limit: true
    }
  }
}
```

```

    },
    assessment: {
      category: "Partially Addressed",
      explanation: "The student partially addressed the task by
mentioning two valid differences, but exceeded the maximum allowed
errors."
    }
  }
}

```

En este caso, al igual que el anterior, el modelo fallo en la etapa final de la evaluación. De todas formas, es interesante ver cómo clasificó las diferencias correctamente por más de que estas fueran bastante vagas en su explicación. Además, se destaca como, si bien fue permisivo, no consideró como válida la sentencia “Thing in room is different place. Everything look not same.” por no referirse a algo específico, como sí lo hacían las otras.

Conclusiones del Análisis El modelo muestra una buena capacidad para evaluar las respuestas de los estudiantes, especialmente en la identificación de diferencias válidas y la tolerancia a errores gramaticales o uso de lenguaje simple. No obstante, se observan inconsistencias en la aplicación de los criterios de evaluación, especialmente en la clasificación final de las respuestas. De todas formas, los experimentos realizados muestran la utilidad de esta herramienta para evaluar una tarea compleja donde hay que analizar muchos componentes a la vez. Podría ser útil intentar quitar algunas responsabilidades al modelo, quizás analizando las diferencias individualmente, y luego calculando “por fuera” el puntaje final en base a las diferencias halladas.

4.8.5. Análisis de Consistencia

Esta subsección se centra en el análisis de consistencia para los tres *prompts* que mostraron mejor desempeño para las tareas de tipo 1: el basado en pensamiento paso a paso (*chain of thought*), el que emplea aprendizaje de pocos ejemplos (*few-shot*), y la combinación de ambos enfoques. Para cada uno se realizaron 30 ejecuciones independientes sobre el mismo conjunto de datos, permitiendo obtener intervalos de confianza y analizar la variabilidad en las predicciones.

El intervalo de confianza del 95 % se calculó utilizando la aproximación normal, apropiado dado el tamaño de muestra de 30 ejecuciones. Además, se calculó la tasa de consistencia promedio para cada prompt (**ConsistencyRatio**), que mide la estabilidad promedio de las predicciones. Para cada entrada de la base de datos, se calcula la frecuencia relativa de la predicción más común (número de veces que aparece la predicción más frecuente dividido entre las 30 ejecuciones). El promedio de estas frecuencias sobre todas las entradas genera un valor entre 0 y 1, donde valores más cercanos a 1 indican mayor estabilidad en las predicciones.

Como métrica complementaria, se evaluó la concordancia entre las diferentes ejecuciones del modelo calculando el coeficiente *Weighted Kappa* tomando pares de ejecuciones (es decir, $\binom{30}{2} = 435$ comparaciones). Mientras que la tasa de consistencia mide la estabilidad a nivel de entrada individual, este análisis permite evaluar qué tan consistente es el modelo consigo mismo entre pares de evaluaciones, independientemente de su acuerdo con las etiquetas reales. La media y desviación estándar de estos coeficientes *Kappa* proveen una medida de la estabilidad interna del modelo. Un *Kappa* medio (referido como **PairwiseKappa**), cercano a 1, con baja desviación estándar, sugiere que el modelo mantiene criterios consistentes a través de diferentes ejecuciones.

En la tabla 4.8 se recopilan los resultados.

Prompt reference	F1	IC 95 %	Kappa	IC 95 %	ConsistencyRatio	PairwiseKappa
cot [7]	0.7123	[0.7061, 0.7184]	0.7698	[0.7636, 0.7760]	0.8592	0.8443
few_shot [9]	0.7199	[0.7169, 0.7229]	0.8105	[0.8066, 0.8145]	0.9575	0.9621
Combinado [10]	0.7534	[0.7481, 0.7588]	0.8251	[0.8210, 0.8291]	0.9099	0.9100

Tabla 4.8: Métricas por *prompt* en 30 ejecuciones

Análisis por Tipo de *Prompt*

El *prompt* *cot* mostró una variabilidad moderada en sus predicciones en términos de la desviación estándar para el F1 y el coeficiente *Kappa*, de 0.0162 y 0.0163 respectivamente. Se obtuvo una tasa de consistencia de 0.85, identificando dos muestras con alta variabilidad y 110 casos (de 286) con predicciones perfectamente consistentes entre todas las evaluaciones.

El análisis de concordancia entre ejecuciones arrojó un *Kappa* medio de 0.8443 (std: 0.0181), indicando un buen nivel de acuerdo interno pero con margen de mejora.

El enfoque **few_shot** demostró una estabilidad mayor, con desviaciones estándar de 0.0080 para F1 y 0.0104 para *Kappa*. No se encontraron muestras altamente variables, y 234 casos mantuvieron predicciones perfectamente consistentes. La concordancia entre ejecuciones mejoró significativamente, alcanzando un *Kappa* medio de 0.9621 (std: 0.0068), lo que sugiere que proporcionar ejemplos ayuda al modelo a establecer un criterio de evaluación más definido.

El *prompt combinado* logró potenciar las fortalezas de ambos enfoques. Si bien su tasa de consistencia (0.9099) fue ligeramente inferior al del **few_shot** puro, sus métricas de desempeño fueron superiores en términos de *Weighted Kappa*. Respecto a la estabilidad, obtuvo 164 casos con consistencia perfecta y un *Kappa* medio entre ejecuciones de 0.9100 (std: 0.0123).

En la figura A.2, presente en el anexo, se encuentra una gráfica violín que permite ver como se distribuyen los valores de *Weighted Kappa* para pares de ejecuciones. Se observa que en la gran mayoría de los casos hubo alta concordancia. Notar como el *prompt few_shot* acumula los valores más cercanos, lo cual muestra estabilidad en la concordancia entre los pares de ejecuciones. Por otro lado, el basado en *cot* se observa un poco más disperso.

Por otra parte, la figura A.3, también presente en el anexo, muestra el valor de la tasa de consistencia por muestra. Aquí también es claro como el *prompt* de mayor inestabilidad es el de *chain of thought*, el más estable el de *few shot*, y con una estabilidad intermedia el enfoque *combinado*.

Observaciones Generales

Los resultados revelan patrones interesantes sobre el comportamiento del modelo bajo las diferentes estrategias de *prompting* utilizadas.

La incorporación de ejemplos (*few shot*) redujo significativamente la variabilidad en las predicciones, probablemente debido a la inclusión de referencias concretas que ayudan a establecer un criterio de evaluación más claro.

El razonamiento paso a paso (*chain of thought*), aunque efectivo para mejorar el nivel de las evaluaciones, mostró mayor variabilidad que el caso con ejemplos. Esto sugiere que el modelo puede seguir diferentes caminos de razonamiento en cada ejecución, y al no tener lineamientos claros que lo orienten, tiende a cambiar su criterio entre evaluaciones.

La combinación de ambas técnicas resultó en un equilibrio favorable, ya que logró una mejor consistencia entre evaluaciones en comparación a solo utilizar *chain of thought* a la vez que mejoró las métricas de desempeño. Esto de cierta manera confirma lo antes dicho, resaltando la falta de una guía clara en el *prompt* de *chain of thought* utilizado, el cual no le permite mantener un criterio consistente para todos los casos. Dar ejemplos solventa este problema en cierta medida, permitiéndole al modelo discernir y ser más consistente en su evaluación.

Esta evidencia sugiere que, para tareas de evaluación donde la consistencia es tan importante como la precisión, el uso de ejemplos en el *prompt* es un recurso útil, ya sea solo o en combinación con técnicas de razonamiento estructurado.

4.9. Conclusiones del Capítulo

En este capítulo, exploramos el diseño y la evaluación de un componente capaz de evaluar el cumplimiento de tareas en respuestas orales. Definimos un enfoque basado en el uso de herramientas de Reconocimiento Automático del Habla y modelos de lenguaje a gran escala con técnicas de *prompting* para transcribir respuestas orales y analizarlas semánticamente en el contexto de diferentes tareas.

Para llevar a cabo el estudio y validar los diferentes *prompts* en las distintas tareas, construimos una base de datos pequeña. Esta nos permitió realizar pruebas detalladas y evaluar el desempeño del modelo en cada tipo de tarea. El estudio se centró en cuatro tipos de tareas, cada una con desafíos específicos:

Para las **tareas de tipo 1**, que consisten en preguntas simples con respuestas breves, encontramos que las técnicas de *few-shot learning* y *chain of thought* mejoraron significativamente el rendimiento del modelo. En particular, la combinación de ambas técnicas produjo los mejores resultados, evidenciando que proporcionar ejemplos claros y guiar el razonamiento del modelo conduce a evaluaciones más precisas y consistentes. Observamos que pequeños cambios en los *prompts* pueden tener un impacto notable en el rendimiento, lo que destaca la importancia de un diseño cuidadoso y detallado.

En las **tareas de tipo 2**, que involucran preguntas basadas en imágenes, utilizamos descripciones detalladas de las imágenes para compensar la falta de capacidad visual del modelo. Los *prompts* que incorporaron *few-shot learning* demostraron ser especialmente efectivos, logrando altos niveles de acuerdo con las evaluaciones de referencia. Sin embargo, identificamos que la categoría intermedia (*Partially Addressed*) representa un desafío, ya que el modelo muestra dificultades para distinguir entre respuestas parcialmente correctas y completamente correctas. Esto sugiere que se requiere un refinamiento adicional en los *prompts* para manejar mejor estos casos.

Para las **tareas de tipo 3**, que requieren respuestas elaboradas que abordan criterios específicos, implementamos un enfoque de *chain of thought* con análisis estructurado. El modelo demostró una buena capacidad para identificar criterios cumplidos explícita e implícitamente y para evaluar la coherencia global de las respuestas. No obstante, observamos limitaciones en la interpretación de criterios implícitos y cierta inconsistencia en la aplicación de las reglas de evaluación. Estos hallazgos indican que, aunque el enfoque es prometedor, es necesario mejorar la uniformidad y flexibilidad del modelo para reconocer implicaciones contextuales evidentes.

En cuanto a las **tareas de tipo 4**, que consisten en identificar diferencias entre dos imágenes, desarrollamos un enfoque que permite evaluar las respuestas comparándolas con un conjunto finito de diferencias válidas. Aunque el análisis

cuantitativo fue limitado debido al tamaño reducido del conjunto de datos, los ejemplos analizados muestran buena capacidad del modelo para detectar en las respuestas las diferencias mencionadas, comparándolas con la lista de referencias y extrayendo evidencias de las mismas. Sin embargo, es necesario ampliar el conjunto de datos y realizar más pruebas para validar la eficacia del enfoque ante una gran variedad de formas de indicar diferencias.

A lo largo del proceso, enfrentamos desafíos relacionados con la naturaleza estocástica de los LLMs y la importancia de diseñar *prompts* efectivos. La iteración constante y la experimentación fueron clave para ajustar las instrucciones y mejorar el rendimiento del modelo. Además, observamos que la incorporación de ejemplos concretos en los *prompts* ayuda al modelo a comprender mejor los criterios de evaluación.

4.9.1. Líneas a Explorar

Este trabajo abre varias líneas de investigación y desarrollo que podrían mejorar y ampliar los resultados obtenidos:

Ampliación y etiquetado del conjunto de datos: Como trabajo futuro, sería interesante aumentar la cantidad y diversidad de datos, implementando un sistema de etiquetado multinivel. Este no solo consideraría el *task fulfillment*, sino que incorporaría etiquetas adicionales que enriquezcan el análisis. Por ejemplo, la inclusión de etiquetas sobre errores gramaticales (que podrían desglosarse en una matriz más detallada, incluyendo repetición de palabras, conjugaciones erróneas, concordancia, entre otros), así como otros tipos de anotaciones. Este etiquetado enriquecido permitiría no solo evaluar el rendimiento del sistema, sino también realizar análisis estadísticos detallados para identificar qué características o patrones en las respuestas tienen mayor impacto en el desempeño del modelo.

Es fundamental que la ampliación de datos ocurra a través ejemplos con **contexto educativo real**. Esto proporcionaría información sobre la eficacia del sistema y su usabilidad. Por otro lado, es necesaria la retroalimentación de docentes y estudiantes para validar o ajustar el sistema, con el fin de satisfacer mejor las necesidades del entorno educativo.

Refinamiento de los prompts: A partir de la experimentación realizada inferimos que es posible mejorar los prompts para abordar las limitaciones identificadas, como la interpretación de criterios implícitos y la clasificación de respuestas parcialmente correctas. Explorar nuevas formulaciones de prompts y arquitecturas de evaluación podría conducir a evaluaciones más precisas. Por ejemplo, la utilización de prompts basados en ejemplos negativos o la aplicación de técnicas de autoevaluación, podría enriquecer la capacidad del modelo para evaluar respuestas de manera más matizada.

Análisis de consistencia: Dada la naturaleza estocástica de los LLMs, es importante realizar análisis de consistencia más profundos, ejecutando múlti-

ples iteraciones y evaluando la variabilidad en las respuestas del modelo. Esto ayudaría a cuantificar la fiabilidad del sistema y a identificar áreas donde se requiere mayor estabilidad.

Incorporación de modelos más avanzados: Considerar el uso de modelos de mayor tamaño podría implicar una mejora significativa en el rendimiento de las evaluaciones, incluso para los prompts ya contruidos. Además, sería interesante explorar el uso de modelos con capacidad de procesar tanto texto como imágenes, ya que estos pueden interpretar directamente el contenido, eliminando la necesidad de descripciones textuales que podrían omitir detalles de interés. Algo aún mejor incluso, podría ser utilizar modelos con capacidad de procesar audio, lo cual permitiría abstraerse del uso de un ASR particular, y podría ser beneficioso para la interpretabilidad de las características acústicas.

Capítulo 5

Módulo Completo

5.1. Introducción

En los capítulos anteriores se exploran en detalle los componentes individuales que conforman el sistema de evaluación del habla en inglés. En este capítulo, se aborda la integración de estos componentes en una plataforma web unificada y funcional. Esta plataforma consolida el trabajo previo, proporcionando una interfaz intuitiva que facilita la prueba y aplicación práctica de las herramientas desarrolladas.

El objetivo principal de este módulo completo es crear un entorno que permita a los usuarios interactuar de manera sencilla con las herramientas descritas en los capítulos previos. Para lograr esto, se optó por una solución basada en las siguientes tecnologías de desarrollo web.

En el *frontend*, se utilizó React, una biblioteca de JavaScript reconocida por su eficiencia y flexibilidad en la creación de interfaces de usuario interactivas.

Para el *backend*, se empleó Python con FastAPI, un framework moderno que facilita la creación de APIs rápidas y escalables. Como sistema de gestión de base de datos, se optó por SQLite, una solución ligera pero potente que se adapta perfectamente a las necesidades de la aplicación.

A lo largo de este capítulo, se explorará la arquitectura del sistema, su implementación y cómo se integran las partes individuales.

Es importante destacar que el rendimiento de este módulo completo está intrínsecamente ligado al desempeño de los componentes individuales presentados en los capítulos anteriores. Sin embargo, la integración de estos elementos en una plataforma es de utilidad para probar las herramientas de manera sencilla.

5.2. Arquitectura del Sistema

5.2.1. Visión General

La arquitectura de nuestro sistema de evaluación del habla está diseñada siguiendo un enfoque modular, lo que permite una integración fluida de los diferentes componentes desarrollados durante este proyecto. El sistema se basa en un modelo cliente-servidor, donde el **frontend** interactúa con el **backend** a través de una API RESTful, y el backend, a su vez, se comunica con diversos módulos especializados para el procesamiento y análisis del audio.

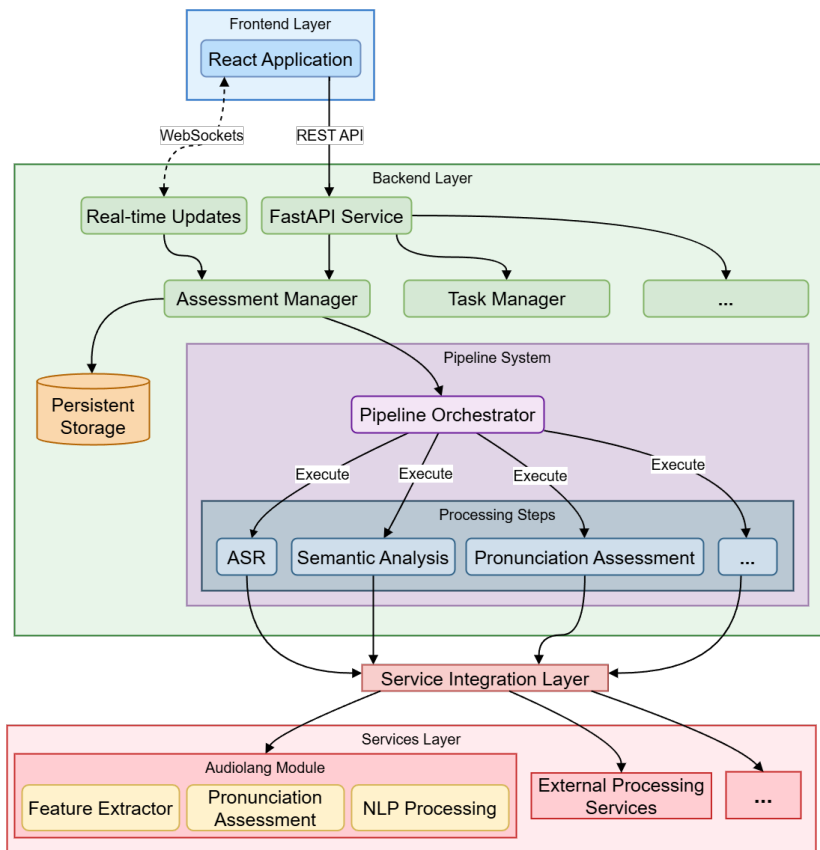


Figura 5.1: Arquitectura general del sistema de evaluación del habla

Como se puede observar en la Figura 5.1, nuestro sistema se compone de tres capas principales:

Capa de Presentación (Frontend). Implementada en React, esta capa proporciona la interfaz con la que interactúa cualquier usuario. Se encarga de la presentación de datos, la captura de audio y la visualización de resultados.

Capa de Lógica de Negocio (Backend). Desarrollada en Python, con

FastAPI, esta capa maneja las solicitudes del cliente, orquesta el flujo de datos entre el frontend y los módulos de procesamiento, y gestiona la persistencia de datos en la base de datos SQLite.

Capa de Servicios (Módulo Audiolang). Esta capa integra al módulo Audiolang, el cual es un módulo Python desarrollado en el marco de este proyecto y se encarga del procesamiento de audio, la extracción de características y la evaluación de los diversos aspectos del habla que hemos abordado en capítulos anteriores. Contiene todos los recursos generados, incluyendo al extractor de características, los métodos de evaluación de características acústicas y las herramientas de evaluación semántica.

Destacar que el proyecto *frontend* y *backend* se encuentran en un mismo repositorio, mientras que *audiolang* se encuentra en un repositorio aparte.

La comunicación entre estas capas se realiza a través de interfaces detalladamente definidas. Esta estructura facilita el mantenimiento y la posibilidad de evolucionar cada componente de forma independiente.

En las siguientes secciones, detallaremos cómo estos módulos se integran para formar el sistema descrito.

5.2.2. Capa de Presentación (frontend)

El frontend de la aplicación se desarrolló en React, organizado en torno a diferentes páginas y componentes que implementan las funcionalidades del sistema.

En cuanto a la **selección y gestión de usuarios**, se dispone de la **UserSelectionPage**, que permite a los usuarios existentes acceder a la aplicación, y la **CreateUserPage**, que facilita el registro de nuevos usuarios en el sistema. Cabe destacar que este manejo de usuarios tiene como único propósito organizar y agrupar las evaluaciones realizadas, sin implementar servicios de autenticación o seguridad adicionales. Además, no existe diferenciación de roles y/o permisos, por lo que cada usuario tiene acceso a la totalidad de las funcionalidades del sistema.

Para la **creación y gestión de tareas**, se realiza una página que se llama **CreateTaskPage**, la cual permite crear nuevas tareas de evaluación con campos como nombre, descripción y requisitos, y la **TaskListPage**, que muestra una lista de tareas disponibles para el usuario.

Los tipos de tareas disponibles corresponden a los cuatro tipos definidos en la Sección 2.3. Estos son incluidos en la categoría de *respuesta abierta*, en la cual se evalúa el *task fulfillment*, tal como se presentó en el Capítulo 4.1. Además, se analizan características acústicas, como se explicó en el Capítulo 3, utilizando la transcripción generada por el ASR como texto de referencia. Esto puede introducir errores si la transcripción no se ajusta a lo que realmente fue dicho. Considerando esto, se define una categoría de *respuesta cerrada* donde se evalúan únicamente las características acústicas. En este caso, la tarea consiste en que el evaluado lea un texto provisto, y las métricas se calculan comparando

el texto dado con la lectura realizada.

En la sección de **evaluación de tareas**, la **TaskAssessmentPage** permite a los usuarios seleccionar una tarea y grabar su respuesta de audio. La grabación y visualización en tiempo real del audio se manejan mediante el componente **AudioRecorder**, que utiliza la API de **MediaRecorder**. Además, para mejorar la experiencia del usuario, la reproducción del audio se maneja con la librería **WaveSurfer.js**, proporcionando visualizaciones interactivas de las formas de onda. Esto permite a los usuarios revisar su grabación antes de enviarla para evaluación. Posteriormente, una vez iniciado el proceso de evaluación, la **TaskAssessmentPage** muestra un indicador de progreso en tiempo real que detalla la etapa actual del procesamiento y proporciona información sobre los modelos que se están ejecutando. En caso de error, el sistema muestra una advertencia, permitiendo al usuario regresar a la etapa de envío y reenviar su grabación. En el *backend*, el error es registrado junto con los detalles necesarios para facilitar el análisis y la identificación de la causa del fallo.

La **visualización de resultados y su interpretación** varía según el tipo de tarea evaluada. Para las tareas de *respuesta abierta*, se muestran la transcripción del audio, el resultado del **task_fulfillment**, y, en caso de que el prompt seleccionado incluya un **chain of thought**, se presentan los pasos de razonamiento asociados a la clasificación. Adicionalmente, se agrega el resultado del análisis de las características acústicas. Para las tareas de *respuesta cerrada*, solo se muestra el resultado del análisis de características acústicas. Un ejemplo de dichas salidas se puede ver en la figura 5.2.

Para la **visualización de resultados e historial**, implementamos la **AssessmentResultsPage**, que muestra los resultados detallados de una evaluación, y la **AssessmentHistoryPage**, que permite consultar el historial de evaluaciones realizadas por el usuario.

La navegación entre estas páginas se gestiona mediante React Router, que proporciona un enrutamiento declarativo. Cada página de la aplicación está compuesta por componentes React reutilizables, lo que mejora la mantenibilidad del código y permite una mayor consistencia en la interfaz de usuario. Algunos de los componentes clave incluyen el **Header**, que proporciona navegación consistente en toda la aplicación, y el **Card.js**, que enmarca el contenido en tarjetas para su mejor visualización.

La interacción con el **backend** se realiza utilizando la **Fetch API**, encapsulando las llamadas en un archivo de servicios (*apiServices.js*) que facilita su reutilización. Además, se implementa comunicación en tiempo real mediante **WebSockets** permitiendo a los usuarios recibir actualizaciones instantáneas sobre el progreso de las evaluaciones.

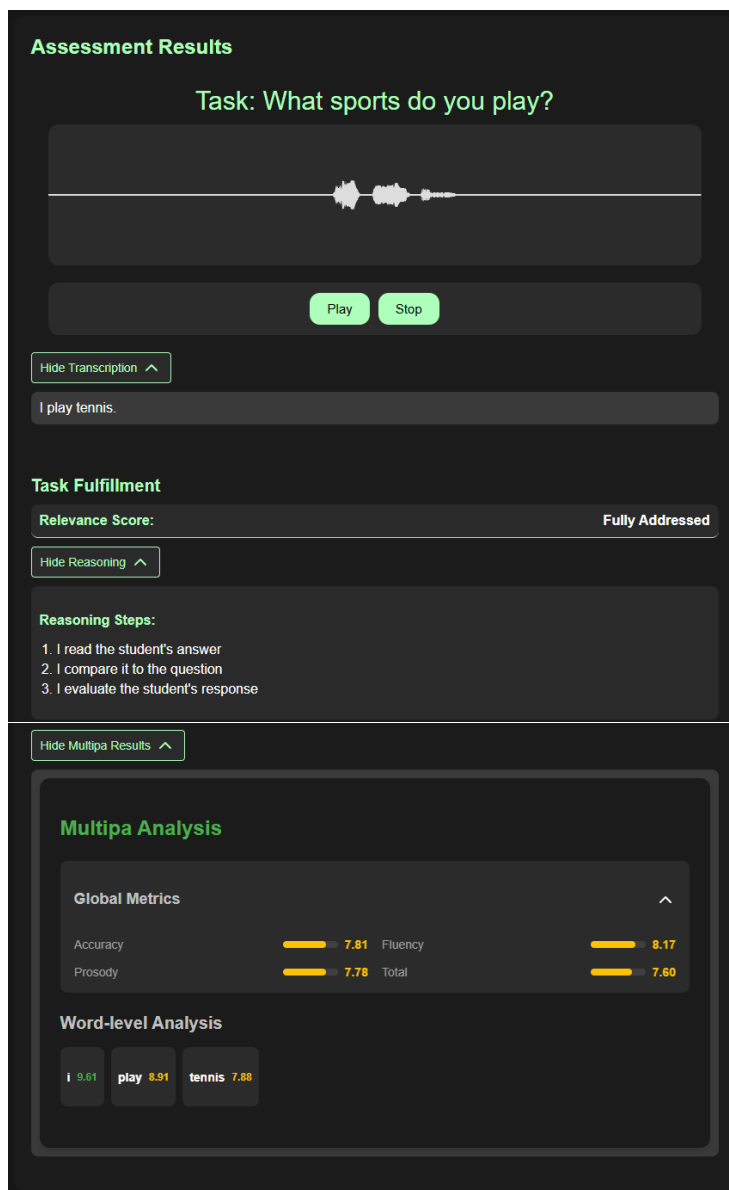


Figura 5.2: Capturas extraídas de la web para una tarea de tipo 1. Se observa transcripción generada, clasificación y razonamiento de analizador semántico y evaluación de características acústicas mediante *MultiPA*.

5.2.3. Capa de Lógica de Negocio (backend)

El *backend* del sistema se implementó utilizando FastAPI, un framework web para Python reconocido por su facilidad de uso. Esta elección permitió crear una API basada en REST que funciona como un puente entre el *frontend* y los módulos de procesamiento de audio, orquestando las evaluaciones y exponiendo servicios para la interacción con la interfaz de usuario. El sistema está organizado en módulos bien definidos, cada uno responsable de una parte específica del proceso, como la gestión de la base de datos, la definición de modelos, la validación de datos, los servicios de evaluación de audio y el manejo de rutas. Esta organización asegura una clara separación de responsabilidades, lo que contribuye a un desarrollo más estructurado y mantenible.

Dentro de las funcionalidades clave que se exponen a través de la API se encuentran:

La **gestión de usuarios**, con endpoints que permiten crear nuevos usuarios y recuperar listas de usuarios existentes.

La **gestión de tareas**, que incluye endpoints para la creación y recuperación de tareas de evaluación.

La **evaluación de tareas**, mediante un endpoint que permite a los usuarios subir archivos de audio para la evaluación de tareas específicas de manera asíncrona. Además, se implementan conexiones por **WebSockets** que facilitan actualizaciones en tiempo real, permitiendo a los usuarios recibir notificaciones instantáneas sobre el progreso de sus evaluaciones.

La **gestión de evaluaciones**, que proporciona endpoints para recuperar todas las evaluaciones asociadas a un usuario.

Para implementar las evaluaciones, se integra el módulo **Audiolang** como un submódulo de Git dentro del directorio de *servicios*. Esta integración permite versionar el código de procesamiento de audio de forma independiente, manteniendo una clara separación entre los diferentes componentes del sistema. **Audiolang** incluye herramientas para la evaluación de los audios según los distintos tipos de tarea, permitiendo la configuración de componentes individuales, como el ASR, tanto para su ejecución local como mediante servicios remotos. Su diseño admite la extensibilidad, lo que permite integrar nuevos modelos sin alterar la estructura base.

Para configurar diferentes flujos de evaluación y seleccionar las herramientas con sus respectivos parámetros, se ha implementado un **patrón de pipeline**. Este patrón permite definir de manera declarativa los pasos que se siguen en el procesamiento de evaluaciones, especificando los modelos a utilizar en cada etapa y sus configuraciones, tales como el entorno de hospedaje (ver Listado 5.1). Los **pipelines** son configurables dentro de un conjunto limitado de implementaciones disponibles para cada componente; sin embargo, la extensibilidad de estos permite integrar nuevas implementaciones en el flujo de evaluación sin costos adicionales significativos. Por ejemplo, el módulo de análisis semántico se podría extender para utilizar un modelo LLM diferente o para ejecutar el

modelo en otro backend, como *VLLM* en lugar del backend de *Hugging Face*, ajustando solo las configuraciones necesarias sin modificar el flujo general de evaluación.

Además, como se muestra en el Listado 5.1, es posible ejecutar múltiples pipelines de evaluación en un mismo ejercicio (como en el ejemplo, donde se realizan evaluaciones tanto de *MultiPA* como de *análisis semántico*). Si estos pipelines comparten pasos idénticos, el sistema evita repeticiones al reutilizar los resultados obtenidos en la primera ejecución de dichos pasos, evitando, por ejemplo, una transcripción duplicada cuando ambos pipelines requieren transcribir el audio con la misma herramienta.

A futuro, se podría extender la definición de pipelines para admitir bifurcaciones o configuraciones más complejas, aunque en el contexto actual y dado el número limitado de pasos en cada pipeline, no se consideró prioritario. También se podría explorar la paralización de algunos pasos; sin embargo, esta funcionalidad no fue implementada debido a las limitaciones de recursos computacionales, especialmente al correr en hardware local.

Para ilustrar lo antes dicho, el siguiente ejemplo configura un pipeline para una tarea de respuesta abierta con pasos de *análisis semántico* y *MultiPA*:

Listing 5.1: Ejemplo de definición de pipelines para tareas de respuesta abierta

```
[
  {
    "name": "task_assessment_pipeline",
    "steps": [
      {
        "name": "asr_server",
        "model": "WhisperX",
        "config": {"model": "http://localhost:4000"}
      },
      {
        "name": "semantic_analysis",
        "model": "Groq",
        "config": {"model": "llama3-8b-8192"}
      }
    ]
  },
  {
    "name": "multipa_pipeline",
    "steps": [
      {
        "name": "asr_server",
        "model": "WhisperX",
        "config": {"model": "http://localhost:4000"}
      },
      {
        "name": "multipa",
        "model": "multipa",
        "config": {"model": "http://localhost:4000"}
      }
    ]
  }
]
```

Otro ejemplo que configura un pipeline para una tarea de *respuesta cerrada* incluye pasos de extracción de fonemas y comparación:

Listing 5.2: Ejemplo de pipeline para tarea de respuesta cerrada

```
[
  {
    "name": "close_response_full_pipeline",
    "steps": [
      {
        "name": "audio_phoneme_extraction",
        "model": "wav2vec2phoneme",
        "config": {"model": "facebook/wav2vec2-lv-60-espeak-cv-ft"}
      },
      {
        "name": "text_phoneme_extraction",
        "model": "phonemizer",
        "config": {"model": "espeak"}
      },
      {
        "name": "phoneme_scoring",
        "model": "custom",
        "config": {"model": "custom"}
      }
    ]
  },
  {
    "name": "main_pipeline",
    "steps": [
      {
        "name": "multipa",
        "model": "multipa",
        "config": {"model": "http://192.168.1.100:8000"}
      }
    ]
  }
]
```

Para el almacenamiento de datos, se integra SQLite con SQLAlchemy como *Object-Relational Mapping* (ORM), lo cual facilita una interacción eficiente con la base de datos al abstraer los detalles de bajo nivel. Elegimos SQLite por ser una solución ligera, suficiente para las necesidades que tenemos en el proyecto.

En la siguiente sección, se describe con mayor detalle el módulo Audiolang, explicando cómo se integra con el backend y contribuye a la funcionalidad general del sistema.

5.2.4. Módulo Audiolang

El módulo **Audiolang** constituye un componente fundamental del sistema, ya que integra todo lo desarrollado en los capítulos anteriores en relación con el procesamiento de audio y la evaluación de diversos aspectos del habla. Este módulo se implementó como un proyecto de Python independiente, lo que facilita su reutilización en otros proyectos. Está compuesto por varios componentes, que se describen a continuación:

Feature Extractor

El **Feature Extractor** es un componente clave del módulo Audiolang, implementado para extraer múltiples características a partir de audios de manera flexible. Este framework está estructurado de forma modular, permitiendo integrar diversos métodos y herramientas de procesamiento de audio. A continuación, se describen sus componentes y funcionamiento.

- **Gestor de Características (Feature Manager):** Coordina el proceso completo de extracción. Gestiona las dependencias entre los extractores y organiza el flujo de procesamiento.
- **Extractores de Características:** Son módulos independientes que extraen información específica del audio, como transcripciones, fonemas o alineaciones. Pueden operar de forma aislada o en conjunto con otros extractores.
- **Backends de Procesamiento:** Proporcionan las implementaciones concretas para las tareas de procesamiento, utilizando modelos o bibliotecas especializadas.

El funcionamiento general se ilustra de manera simplificada en la Figura 5.3. A continuación se detalla cada paso del proceso:

1. El usuario define una configuración que especifica las características a extraer y establece las dependencias entre los extractores.
2. El Gestor de Características utiliza esta configuración para construir un grafo de dependencias y determinar el orden de ejecución de los extractores.
3. Se procesan los audios en lotes (batch processing) por cada extractor.
4. Cada extractor recibe como entrada el batch completo de audios y los resultados de los extractores de los que depende. Los extractores basados en modelos siguen un ciclo de carga/procesamiento/liberación del modelo para el batch completo, mientras que los extractores simples solo realizan la extracción directa sobre el batch.

5. Los resultados generados por cada extractor se incorporan al proceso, quedando disponibles para los siguientes extractores según el orden de ejecución determinado.
6. Se retornan todas las características extraídas para el batch de audios procesado.

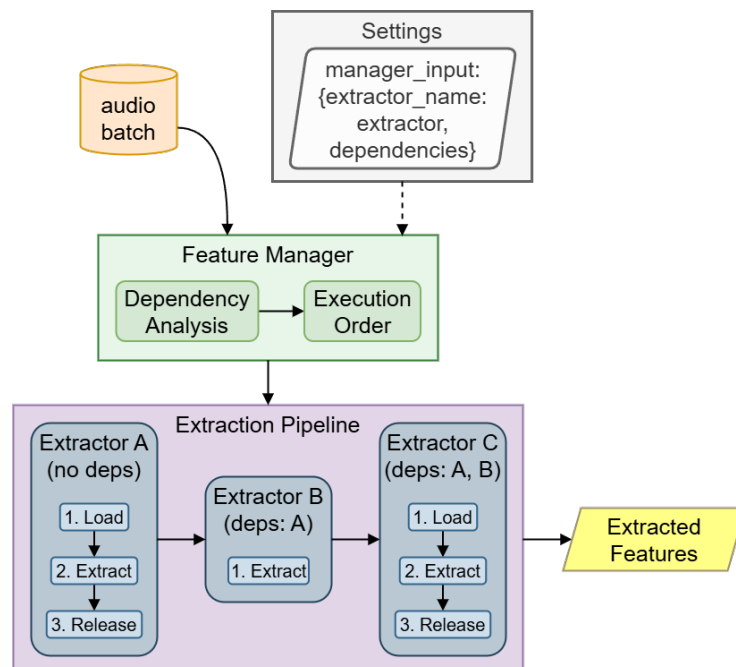


Figura 5.3: Flujo simplificado del Feature Extractor.

La configuración del *Feature Extractor* se realiza mediante un diccionario que declara los extractores y sus dependencias. Un ejemplo de configuración se muestra en el Listado 5.3.

Listing 5.3: Ejemplo de configuración del *Feature Extractor*.

```

manager_input = {
    "speech_to_text": {
        "extractor": speech_to_text_extractor,
        "dependencies": []
    },
    "aligned_text": {
        "extractor": aligned_text_extractor,
        "dependencies": ["speech_to_text"]
    },
    "simple_features": {

```

```

        "extractor": simple_feature_extractor,
        "dependencias": ["aligned_text"]
    }
}

```

En este ejemplo, el extractor `aligned_text` depende de `speech_to_text`, asegurando que las transcripciones estén disponibles antes de realizar la alineación. Este enfoque permite al sistema gestionar automáticamente el orden de ejecución y las dependencias entre extractores.

Este framework sigue los siguientes principios de diseño:

Modularidad. Cada extractor de características es un componente independiente que puede ser utilizado de forma aislada o en combinación con otros. Los componentes están bien separados, lo que facilita el mantenimiento y la comprensión del código.

Extensibilidad. Gracias al uso de clases abstractas y patrones de diseño como el *Patrón Strategy* y la *Inyección de Dependencias*, es sencillo agregar nuevos extractores de características o implementaciones para extractores ya existentes sin tener que modificar el código existente.

Escalabilidad. Se construye el sistema para procesar un conjunto de audios, implementando una estrategia de extracción por lotes (*batch processing*). Se implementa un sistema de caché para almacenar temporalmente los datos de audio cargados, reduciendo los accesos a disco y mejorando la eficiencia. El procesamiento por lotes se realiza de la siguiente manera: el lote completo de datos se carga y se mantiene en memoria una sola vez, evitando lecturas repetitivas de disco. Luego, cada extractor de características se carga en memoria, procesa el lote completo y luego se libera antes de pasar al siguiente extractor. Esto minimiza la carga y descarga repetitiva, mejora el rendimiento y permite un uso eficiente de la memoria al procesar grandes volúmenes de datos. Esto es especialmente útil cuando el extractor es implementado utilizando modelos que requieren gran cantidad de memoria. Al cachear el lote en memoria, reducimos significativamente el tiempo de acceso a los datos, lo cual es especialmente beneficioso cuando trabajamos con conjuntos de datos extensos.

Configurabilidad. Ofrece opciones de configuración flexibles para adaptar el comportamiento de los extractores a diferentes escenarios de uso. La configuración del *Feature Extractor* se realiza de manera declarativa mediante un diccionario que especifica los extractores y sus dependencias, como se muestra en el Listado 5.3. Esto facilita la adición de nuevos extractores y la gestión de sus dependencias sin necesidad de modificar el código fuente. Además, el sistema incorpora mecanismos de validación de la configuración para detectar dependencias no resueltas o ciclos, y maneja los errores previo a procesar.

Un ejemplo de uso del *Feature Extractor* se presenta en el Listado 5.4.

Listing 5.4: Ejemplo de uso del *Feature Extractor*.

```

# Inicialización del FeatureManager
feature_manager = FeatureManager(manager_input=manager_input)

```

```

# Lista de rutas a archivos de audio
audio_batch = [
    "path/to/audio1.wav",
    "path/to/audio2.wav",
    "path/to/audio3.wav"
]

# Lista de transcripciones (vacía si no hay)
transcriptions = [" " for _ in audio_batch]

# Extracción de características
extracted_features = feature_manager.extract_features(batch=
    audio_batch, transcriptions=transcriptions)

# Visualización de los resultados
for i, features in enumerate(extracted_features):
    print(f"Características para {audio_batch[i]}:")
    for feature_name, value in features.items():
        print(f"    {feature_name}: {value}")
    print("\n")

```

Este enfoque permite extraer características de manera eficiente y estructurada, facilitando su uso en sistemas de evaluación o análisis posteriores.

Además del Feature Extractor, Audiolang incluye todas las implementaciones que desarrollamos en los capítulos anteriores. Principalmente, está compuesto por varios sistemas de evaluación que utilizan las características extraídas.

Evaluación de Pronunciación

Uno de los sistemas construidos sobre el *Feature Extractor* es la *línea base* de evaluación de pronunciación presentada en el Capítulo 3. Esta hace uso del extractor para obtener la secuencia de fonemas de los audios y de los textos de referencia, las cuales luego compara para devolver métricas de pronunciación.

Adicionalmente, se integra el submódulo **MultiPA** para proporcionar herramientas adicionales en el análisis detallado de las características del audio. Para lograr esta integración, adaptamos la API de inferencia original de MultiPA, que estaba diseñada para utilizar un texto generado por su propio ASR, de modo que permita introducir un texto de referencia externo como entrada para la evaluación. Esto permite evaluaciones más precisas, ya que el ASR usado por *MultiPA* (Whisper) es un poco menos preciso que el que seleccionamos luego del análisis de ASR (WhisperX). Asimismo, para los ejercicios en los que el evaluado lee un texto fijo, es ideal pasar dicho texto directamente como referencia.

Procesamiento del Lenguaje Natural (NLP)

Gracias a la *abstracción* y *polimorfismo* implementados, es posible intercambiar entre diferentes modelos sin necesidad de modificar el código que los utiliza.

Esto se logra mediante clases base que definen interfaces comunes, permitiendo que las implementaciones específicas de cada modelo se integren de forma transparente. Un elemento clave de este diseño es la clase `ConversationModel`, que sirve como base para los modelos que manejan conversaciones estructuradas.

La clase `ConversationModel` define métodos esenciales para manejar interacciones que incluyen roles como *sistema*, *usuario* y *asistente*. Esto permite construir y gestionar conversaciones paso a paso, manteniendo el contexto y facilitando la generación de respuestas coherentes. Los mensajes se formatean según plantillas (*prompts*) que guían al modelo en la evaluación de aspectos específicos de las respuestas de los usuarios.

Este enfoque permite, por ejemplo, analizar la coherencia y relevancia de las respuestas en tareas de comprensión lectora o conversación, adaptando fácilmente el sistema a diferentes escenarios. Además, el módulo NLP está diseñado para procesar las respuestas en lotes (*batch processing*), optimizando el rendimiento y permitiendo el análisis de grandes volúmenes de datos.

El código desarrollado incluye integraciones con varias implementaciones del modelo, como *vLLM* y *Hugging Face* para ejecutar Llama, así como con modelos externos como *ChatGPT* y *Claude*, siendo posible extenderlo aún más.

5.3. Flujo al evaluar una tarea

Como se ilustra en la Figura 5.4, el flujo de interacción entre el frontend y el backend para evaluar una tarea es el siguiente:

1. El usuario selecciona una tarea en la aplicación web y graba su respuesta.
2. Al finalizar la grabación, el usuario envía el audio al backend mediante el endpoint `POST /evaluate-task`.
3. El backend guarda el archivo de audio, lo convierte al formato adecuado y crea una entrada en la base de datos para la evaluación en estado Pendiente.
4. Se inicia una tarea asíncrona que procesa la evaluación siguiendo el pipeline configurado para el tipo de tarea.
5. El backend devuelve al frontend un **ID de evaluación**, que sirve para consultar el estado de la evaluación.
6. El frontend se conecta al backend mediante el endpoint `GET /ws/assessment/{assessment_id}` para recibir actualizaciones en tiempo real a través de WebSockets.
7. A medida que se procesan los pasos del pipeline, el backend envía actualizaciones al frontend a través de la conexión WebSocket.
8. Una vez completada la evaluación, el frontend muestra los resultados al usuario.

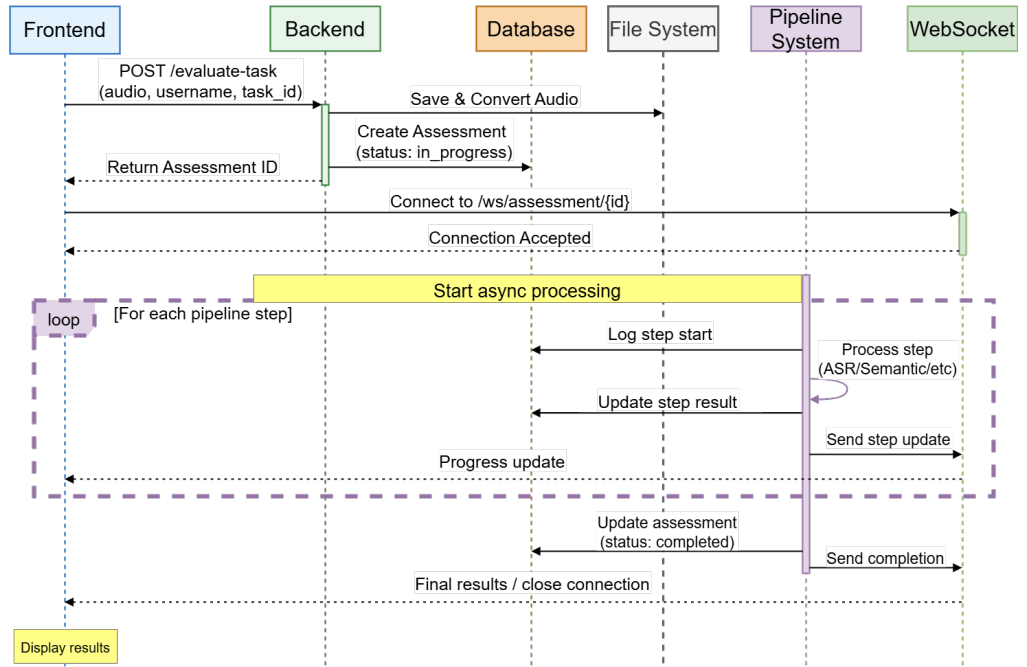


Figura 5.4: Diagrama de flujo simplificado del proceso de evaluación de tareas

Destacar que en caso de que la tarea asíncrona no se completase o existiese un error, esto se envía a través de la conexión WebSocket y se redirige al usuario a la sección de evaluación, con el audio grabado, para que lo pueda reenviar.

5.4. Conclusiones del Capítulo

En este capítulo, se aborda la integración de todos los componentes desarrollados previamente en una plataforma web. Esta integración permite crear un entorno interactivo que facilita la prueba y aplicación práctica de las herramientas de evaluación del habla en inglés que hemos desarrollado a lo largo del proyecto.

La arquitectura modular del sistema, basada en una separación clara entre el *frontend*, el *backend* y el módulo de procesamiento de audio (*Audiolang*), ha demostrado ser efectiva para organizar el flujo de datos y permitir una comunicación fluida entre los distintos componentes. Esta estructura no solo mejora la mantenibilidad del código, sino que también proporciona la flexibilidad necesaria para expandir la plataforma con nuevas herramientas y funcionalidades en el futuro.

La plataforma fue validada tanto en entornos locales como en situaciones más desafiantes, como su presentación en la exposición Ingeniería deMuestra (octubre de 2024) [67]. Para superar las limitaciones de hardware de la laptop

usada en la exposición, se aprovechó la flexibilidad de la arquitectura que permite ejecutar las herramientas tanto local como remotamente. Esto permitió hospedar los procesos más pesados en otra máquina y consumirlos desde la laptop, logrando que la aplicación funcionara correctamente. Destacar que a pesar del ambiente ruidoso en el que se probó, la aplicación tuvo un comportamiento adecuado. Las figuras presentes en 5.5 ilustran la plataforma siendo utilizada en la exposición.



Figura 5.5: Plataforma siendo utilizada en la exposición *Ingeniería deMuestra*, 11 y 12 de octubre de 2024.

La integración de los componentes en una plataforma web unificada ha consolidado el trabajo realizado, resultando en una herramienta funcional cuya arquitectura modular facilita futuras expansiones.

Capítulo 6

Conclusiones y Trabajo Futuro

En este trabajo, se exploraron y desarrollaron herramientas para la **automatización de la evaluación del habla en inglés como segunda lengua**. Partiendo de la necesidad planteada por el programa *Ceibal en Inglés* de evaluar oralmente a un gran número de estudiantes a nivel nacional, la propuesta fue diseñar y evaluar métodos que permitan abordar esta tarea de manera escalable.

Inicialmente, se investigaron las técnicas actuales de evaluación automatizada del habla y su aplicación en el aprendizaje de idiomas. Se analizaron los fundamentos teóricos y los métodos existentes, lo que permitió identificar las áreas en las que enfocarnos.

Para abordar la **evaluación de características acústicas**, tomamos dos enfoques: una línea base basada en la comparación fonética y modelo multitarea llamado *MultiPA* preentrenado sobre una base de datos de acceso público. A través de pruebas y evaluaciones, constatamos que la línea base presentaba limitaciones significativas, mientras que *MultiPA* mostró resultados prometedores en la evaluación de la inteligibilidad del discurso al realizar inferencias sobre la base de datos de Ceibal.

En cuanto a la **evaluación de características semánticas**, se exploró el uso de modelos de lenguaje de gran escala y técnicas de *prompting* para analizar el cumplimiento de tareas en las respuestas orales de los estudiantes. Se diseñaron y evaluaron diversos *prompts* en cuatro tipos de tareas, cada una con desafíos específicos. Para validar estos *prompts*, se construyó una base de datos pequeña que permitió realizar pruebas detalladas y evaluar el desempeño del modelo en cada tipo de tarea. Todas las conclusiones obtenidas se basan en la evaluación realizada sobre estos ejemplos generados.

Las técnicas de **few-shot learning** y **chain of thought** demostraron ser particularmente efectivas en nuestro estudio, donde su implementación en par-

tes o de manera conjunta, dieron los resultados más significativos. Si bien se identificaron algunos desafíos específicos en la categorización de respuestas parcialmente correctas y en la interpretación de criterios implícitos, la estrategia de proporcionar ejemplos claros junto con un razonamiento guiado resultó en evaluaciones más precisas y consistentes.

Finalmente, se integran todos los componentes desarrollados en una **plataforma web unificada**, lo que permite probar las herramientas de manera sencilla y funcional. Esta plataforma acerca las soluciones desarrolladas a un contexto práctico, lo cual evidencia la aplicabilidad de las herramientas construidas en contextos reales.

Entre las principales **contribuciones** de este trabajo, se destacan:

- La implementación y evaluación de métodos para la evaluación automática de características acústicas, enfocado en la validación de resultados para el público objetivo.
- La implementación y evaluación de un componente capaz de evaluar el *task fulfillment* a partir de respuestas orales.
- La integración de los componentes desarrollados en una plataforma web unificada, con el fin de facilitar su uso y evaluación en contextos prácticos.

Sin embargo, se enfrentaron **dificultades y limitaciones** a lo largo del proyecto. Una de las principales fue la escasez de bases de datos disponibles para construir y evaluar herramientas de evaluación oral. En particular, para el análisis de características acústicas, la ausencia de un conjunto de datos representativo del público objetivo con etiquetas relevantes de pronunciación restringió considerablemente las opciones metodológicas. Esta limitación impactó el desarrollo del proyecto en múltiples niveles: **primero**, no fue posible realizar entrenamiento o *fine-tuning* de modelos especializados para evaluar características específicas de pronunciación; **segundo**, se tuvo que construir una línea base más sencilla conceptualmente; y **tercero**, fue necesario restringir la validación de los resultados a los datos disponibles, cuyas etiquetas no eran las ideales, lo que dificultó la medición precisa del desempeño del sistema.

En cuanto a la evaluación del cumplimiento de tareas, aunque se logró construir una base de datos propia para la validación, las limitaciones en su alcance también restringieron el espectro del análisis. Una base de datos más extensa, validada por múltiples expertos y con un esquema de etiquetado más granular, habría permitido explorar dimensiones adicionales en la evaluación de características semánticas, como la corrección gramatical o la riqueza de vocabulario. Estas características, aunque relevantes para una evaluación integral del dominio del idioma, quedaron fuera del alcance del presente trabajo.

Este estudio abre **varias líneas de investigación y desarrollo** que podrían mejorar y ampliar los resultados obtenidos. Sería de interés construir *bases de datos específicamente diseñadas* para cada aspecto a evaluar.

Para las **características acústicas**, esto implica desarrollar conjuntos de datos que representen adecuadamente a nuestro público objetivo con etiquetas generadas por expertos, en características que permitan medir la inteligibilidad. Una vez construida dicha base, se podrían explorar estrategias de *transfer learning* utilizando *MultiPA* como base o desarrollar modelos que, tomando como entrada las métricas generadas por MultiPA, produzcan evaluaciones adaptadas a nuestro contexto específico.

En cuanto a las **características semánticas**, el énfasis estaría en ampliar la cantidad de datos considerando un espectro mayor de etiquetas para validar el sistema en diversos escenarios pero comunes en el lenguaje hablado. Un conjunto de datos más completo sería útil para evaluar el comportamiento del sistema ante diferentes niveles de respuestas, midiendo así el impacto de estas variaciones en el desempeño del sistema. A su vez, considerar el uso de modelos de mayor tamaño y con capacidad para procesar tanto texto como imágenes podría mejorar significativamente el rendimiento en ciertas tareas, eliminando la necesidad de descripciones textuales detalladas.

También sería interesante aprovechar las capacidades de *diarización* de los sistemas ASR modernos, como el que tiene WhisperX, que permiten identificar y etiquetar automáticamente a diferentes hablantes en una conversación. Esto abriría la posibilidad de evaluar la oralidad de distintas formas, por ejemplo al evaluar diálogos completos, analizando aspectos como la capacidad de interacción, la coherencia en el intercambio de ideas y la fluidez de la conversación en su conjunto.

Bibliografía

- [1] Consejo de Europa, *Marco Común Europeo de Referencia para las Lenguas: Aprendizaje, enseñanza, evaluación*. Cambridge University Press, 2001. Traducción al español publicada en 2002 por el Ministerio de Educación, Cultura y Deporte de España en colaboración con el Instituto Cervantes y Grupo Anaya.
- [2] M. Levy, *Computer-Assisted Language Learning: Context and Conceptualization*. Clarendon paperbacks, Clarendon Press, 1997.
- [3] Duolingo, “Duolingo.” <https://es.duolingo.com/>, 2011. Último acceso: 18 de noviembre de 2024.
- [4] Voxy, “Voxy proficiency assessment,” 2024. La evaluación VPA incluye una sección opcional de habla que permite a los estudiantes enviar grabaciones de audio de sus respuestas a preguntas orales y escritas, las cuales son calificadas por evaluadores certificados. Último acceso: 21 octubre 2024.
- [5] ELSA Speak, “Elsa speak.” <https://elsaspeak.com/en/>, 2016. Último acceso: 18 de noviembre de 2024.
- [6] K. Zechner and K. Evanini, eds., *Automated Speaking Assessment: Using Language Technologies to Score Spontaneous Speech*. Routledge, 1st ed., 2019.
- [7] K. Kyriakopoulos, *Deep Learning for Automatic Assessment and Feedback of Spoken English*. PhD thesis, University of Cambridge, 2020.
- [8] J. Cheng, “Automatic assessment of prosody in high-stakes english tests,” in *Interspeech*, pp. 1589–1592, ISCA, 2011.
- [9] P. Tavakoli and C. Wright, *Measuring Fluency*, p. 43–63. Cambridge Applied Linguistics, Cambridge University Press, 2020.
- [10] G. Bailly, E. Godde, A.-L. Piat-Marchand, and M.-L. Bosse, “Automatic assessment of oral readings of young pupils,” *Speech Communication*, vol. 138, pp. 67–79, 2022.

- [11] N. Khabbazzbashi, F. Nakatsuhara, C. Inoue, G. Kaplan, and A. Green, “The design and validation of an online speaking test for young learners in uruguay: Challenges and innovations,” *International Journal of TESOL Studies*, 2022.
- [12] K. M. Knill, L. Wang, Y. Wang, X. Wu, and M. J. Gales, “Non-native children’s automatic speech recognition: The interspeech 2020 shared task alta systems,” in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, ISCA, 2020.
- [13] S. Wills, Y. Bai, C. Tejedor-García, C. Cucchiarini, and H. Strik, “Automatic speech recognition of non-native child speech for language learning applications,” 2023.
- [14] R. Lado, *Language Testing: The Construction and Use of Foreign Language Tests*. Longman, 1961.
- [15] R. Wardhaugh, “The contrastive analysis hypothesis,” *TESOL Quarterly*, vol. 4, no. 2, pp. 123–130, 1970.
- [16] S. Thirakunkovit, “The notions of language proficiency and language dimensionality,” *PASAA*, vol. 55, 2018.
- [17] M. J. Munro and T. M. Derwing, “Foreign accent, comprehensibility, and intelligibility in the speech of second language learners,” *Language Learning*, vol. 45, no. 1, pp. 73–97, 1995.
- [18] J. M. Levis, “Changing contexts and shifting paradigms in pronunciation teaching,” *TESOL Quarterly*, vol. 39, no. 3, pp. 369–377, 2005.
- [19] O. Kang and R. I. Thomson, “The routledge handbook of contemporary english pronunciation,” 2017.
- [20] T. M. Derwing and M. J. Munro, *Pronunciation Fundamentals: Evidence-based Perspectives for L2 Teaching and Research*. Amsterdam: John Benjamins Publishing Company, 2015.
- [21] R. I. Thomson, T. M. Derwing, and M. J. Munro, “The effectiveness of l2 pronunciation instruction: A narrative review,” *Applied Linguistics*, vol. 36, no. 3, pp. 326–344, 2015.
- [22] S. M. Witt, “Automatic error detection in pronunciation training: where we are and where we need to go,” in *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training*, pp. 1–8, 2012.
- [23] Y. E. Kheir, A. Ali, and S. A. Chowdhury, “Automatic pronunciation assessment - a review,” *arXiv preprint arXiv:2310.13974*, 2023. <https://arxiv.org/pdf/2310.13974.pdf>.

- [24] J. Vidal, L. Ferrer, and L. Brambilla, “Epadb: A database for development of pronunciation assessment systems,” in *Proc. Interspeech 2019*, pp. 589–593, 2019.
- [25] S. M. Witt and S. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.
- [26] S. Kanters, C. Cucchiaroni, and H. Strik, “The goodness of pronunciation algorithm: a detailed performance study,” in *The Netherlands and Department of Linguistics, Radboud University Nijmegen*, 2000.
- [27] J. Vidal, C. Bonomi, M. Sancinetti, and L. Ferrer, “Phone-level pronunciation scoring for spanish speakers learning english using a gop-dnn system,” in *Proc. Interspeech 2021*, pp. 4423–4427, 2021.
- [28] J. Zhang, Z. Zhang, Y. Wang, Z. Yan, Q. Song, Y. Huang, K. Li, D. Povey, and Y. Wang, “speechocean762: An open-source non-native english speech corpus for pronunciation assessment,” 2021.
- [29] Y. Gong, Z. Chen, I.-H. Chu, P. Chang, and J. Glass, “Transformer-based multi-aspect multi-granularity non-native english speaker pronunciation assessment,” *arXiv preprint arXiv:2205.03432*, 2022.
- [30] X. Cao, Z. Fan, T. Svendsen, and G. Salvi, “An analysis of goodness of pronunciation for child speech,” in *Proceedings of Interspeech 2023*, pp. 4613–4617, ISCA, 2023.
- [31] W.-K. Leung, X. Liu, and H. Meng, “Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8132–8136, IEEE, 2019.
- [32] X. Xu, Y. Kang, S. Cao, B. Lin, and L. Ma, “Explore wav2vec 2.0 for mispronunciation detection,” in *Interspeech 2021*, pp. 4428–4432, 2021.
- [33] E. Kim, J.-J. Jeon, H. Seo, and H. Kim, “Automatic pronunciation assessment using self-supervised speech representation learning,” 2022.
- [34] Y.-W. Chen, Z. Yu, and J. Hirschberg, “Multipa: A multi-task speech pronunciation assessment model for open response scenarios,” 2024.
- [35] Y. Kheir, A. Ali, and S. Chowdhury, “Automatic pronunciation assessment - a review,” in *Findings of the Association for Computational Linguistics: EMNLP 2023* (H. Bouamor, J. Pino, and K. Bali, eds.), (Singapore), pp. 8304–8324, Association for Computational Linguistics, Dec. 2023.
- [36] M. Sancinetti, J. Vidal, C. Bonomi, and L. Ferrer, “A transfer learning based approach for pronunciation scoring,” 2023.

- [37] Advanced Research Projects Agency, “Arpabet,” 1970. Último acceso: noviembre 2024.
- [38] International Phonetic Association, “Alfabeto fonético internacional (ipa),” 1886. Último acceso: noviembre 2024.
- [39] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, “L2-arctic: A non-native english speech corpus,” pp. 2783–2787, 09 2018.
- [40] J. Vidal, L. Ferrer, and L. Brambilla, “Epadb: A database for development of pronunciation assessment systems,” pp. 589–593, 09 2019.
- [41] Q. Xu, A. Baevski, and M. Auli, “Simple and effective zero-shot cross-lingual phoneme recognition,” *arXiv preprint arXiv:2109.11680*, 2021.
- [42] M. Bernard and H. Titeux, “Phonemizer: Text to phones transcription for multiple languages in python,” *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021.
- [43] R. Dunn *et al.*, “espeak ng,” <https://github.com/espeak-ng/espeak-ng>, 2015. Accessed: 2024-11-07.
- [44] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” 2021.
- [45] J. Zhu, C. Zhang, and D. Jurgens, “Phone-to-audio alignment without text: A semi-supervised approach,” *CoRR*, vol. abs/2110.03876, 2021.
- [46] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized BERT pretraining approach,” *CoRR*, vol. abs/1907.11692, 2019.
- [47] U. Kuckartz, S. Rädiker, T. Ebert, and J. Schehl, *Statistik: Eine verständliche Einführung*. Wiesbaden: Springer VS, 2 ed., 2013.
- [48] D. Knight, “Comprehensive assessment of spoken language,” in *Encyclopedia of Child Behavior and Development*, pp. 642–644, Springer, Boston, MA, 2011.
- [49] “Google cloud speech-to-text.” <https://cloud.google.com/speech-to-text>, 2017. Accessed: 2024-09-04.
- [50] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” *arXiv preprint arXiv:2212.04356*, 2022.
- [51] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

- [52] M. Bain, J. Huh, T. Han, and A. Zisserman, “Whisperx: Time-accurate speech transcription of long-form audio,” *arXiv preprint arXiv:2303.00747*, 2023.
- [53] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.
- [54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [55] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” *OpenAI Blog*, 2018.
- [56] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [57] OpenAI, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [58] A. Llama Team, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [59] K. Taghipour and H. T. Ng, “Automated essay scoring using large language models,” *arXiv preprint arXiv:2203.12345*, 2022.
- [60] T. Nguyen, A. Tran, and H. Nguyen, “Personalized feedback generation for educational tasks using large language models,” *Proceedings of the 16th International Conference on Educational Data Mining (EDM)*, 2023.
- [61] OpenAI, “Gpt-4 technical report.” <https://openai.com/research/gpt-4>, 2023. Accessed: 2024-11-19.
- [62] G. DeepMind, “Gemini: Multimodal ai model.” <https://ai.googleblog.com/2023/12/introducing-gemini.html>, 2023. Accessed: 2024-11-19.
- [63] Anthropic, “Claude 3: The next generation ai assistant.” <https://www.anthropic.com/index/claude-3>, 2024. Accessed: 2024-11-19.
- [64] M. AI, “Llama 3: Open foundation models.” <https://ai.meta.com/llama3>, 2024. Accessed: 2024-11-19.
- [65] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” *arXiv preprint arXiv:2005.14165*, 2020.

- [66] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou, “Chain-of-thought prompting elicits reasoning in large language models,” *arXiv preprint*, vol. arXiv:2201.11903, 2022.
- [67] “Ingeniería demuestra,” 2024. Ingeniería deMuestra 2024.

Anexo A

Anexo 1

A.1. Estadísticas Descriptivas de la Evaluación de *EpaDB* en *MultiPA*

En esta sección se presentan las estadísticas descriptivas obtenidas a partir de la evaluación de la base de datos *EpaDB* utilizando *MultiPA*. Dado que esta base cuenta con etiquetas por audio, fue posible realizar un análisis directo de los resultados inferidos por el modelo. Las tablas incluyen estadísticas descriptivas de las características evaluadas: Precisión, Fluidez, Prosodia y Puntaje Total, agrupadas según las etiquetas correspondientes de cada audio.

G. de N.	Media	Desv. Estándar	Mediana	Mínimo	Máximo
1	7.710	0.726	7.568	6.788	8.975
2	7.640	0.818	7.684	5.437	9.333
3	8.028	0.790	8.145	5.239	9.653
4	8.481	0.599	8.564	6.104	9.707
5	8.383	0.639	8.480	5.915	9.765

Tabla A.1: Estadísticas descriptivas de la Precisión según el grado de natividad.

G. de N.	Media	Desv. Estándar	Mediana	Mínimo	Máximo
1	8.355	0.718	8.257	7.260	9.617
2	8.318	0.755	8.409	6.203	9.803
3	8.792	0.774	8.905	5.799	10.369
4	9.251	0.649	9.340	6.241	10.573
5	9.225	0.708	9.297	6.475	10.602

Tabla A.2: Estadísticas descriptivas de la Fluidez según el grado de natividad.

G. de N.	Media	Desv. Estándar	Mediana	Mínimo	Máximo
1	7.848	0.712	7.664	6.950	9.264
2	7.784	0.729	7.859	5.559	9.117
3	8.242	0.715	8.343	5.548	9.988
4	8.718	0.628	8.782	5.978	10.119
5	8.721	0.689	8.841	5.822	10.021

Tabla A.3: Estadísticas descriptivas de la Prosodia según el grado de natividad.

G. de N.	Media	Desv. Estándar	Mediana	Mínimo	Máximo
1	7.774	0.765	7.513	6.811	9.286
2	7.737	0.802	7.766	5.594	9.419
3	8.156	0.779	8.280	5.315	9.679
4	8.616	0.613	8.739	6.092	9.804
5	8.527	0.676	8.653	5.624	9.641

Tabla A.4: Estadísticas descriptivas del Puntaje Total según el grado de natividad.

A.2. Estadísticas Descriptivas de la Evaluación de la Base de *Ceibal* en *MultiPA*

En esta sección se presentan las estadísticas descriptivas asociadas a la evaluación de la base de datos de *Ceibal* en *MultiPA*, utilizando tanto la media como la mediana como métricas agregadas. Dado que la base cuenta con un puntaje holístico por estudiante, se calcularon estas agregaciones a partir de los resultados obtenidos por *MultiPA* en los audios individuales. Las tablas detallan las estadísticas para las características evaluadas: Precisión, Fluidez, Prosodia y Puntaje Total, agrupadas por el puntaje global de *Delivery*.

A.2.1. Basadas en la Media

Delivery	Media	Desv. Estándar	Mediana	Mínimo	Máximo
0	5.67	0.59	5.49	4.97	6.52
1	5.98	0.61	6.05	5.24	7.21
2	6.95	0.77	7.35	6.07	7.45
3	6.90	0.58	6.66	6.49	7.89

Tabla A.5: Estadísticas descriptivas de Precisión según Delivery, utilizando la media como métrica agregada.

Delivery	Media	Desv. Estándar	Mediana	Mínimo	Máximo
0	6.28	0.57	6.12	5.62	7.16
1	6.67	0.59	6.78	5.70	7.72
2	7.54	0.81	7.78	6.64	8.20
3	7.45	0.60	7.32	6.91	8.40

Tabla A.6: Estadísticas descriptivas de Fluidez según Delivery, utilizando la media como métrica agregada.

Delivery	Media	Desv. Estándar	Mediana	Mínimo	Máximo
0	6.01	0.53	5.87	5.39	6.79
1	6.33	0.52	6.39	5.47	7.29
2	7.11	0.71	7.39	6.30	7.64
3	7.08	0.58	6.91	6.55	8.03

Tabla A.7: Estadísticas descriptivas de Prosodia según Delivery, utilizando la media como métrica agregada.

Delivery	Media	Desv. Estándar	Mediana	Mínimo	Máximo
0	5.81	0.61	5.55	5.17	6.76
1	6.12	0.63	6.16	5.39	7.47
2	7.10	0.79	7.50	6.20	7.61
3	7.05	0.59	6.72	6.68	8.07

Tabla A.8: Estadísticas descriptivas de Puntaje Total según Delivery, utilizando la media como métrica agregada.

A.2.2. Basadas en la Mediana

Delivery	Media	Desv. Estándar	Mediana	Mínimo	Máximo
0	5.60	0.66	5.51	4.81	6.57
1	5.95	0.77	5.98	4.85	7.54
2	6.91	0.82	7.33	5.96	7.43
3	6.86	0.94	6.72	6.07	8.45

Tabla A.9: Estadísticas descriptivas de la Precisión según Delivery, utilizando la mediana como métrica agregada.

Delivery	Media	Desv. Estándar	Mediana	Mínimo	Máximo
0	6.20	0.59	6.09	5.50	6.97
1	6.77	0.76	6.93	5.71	8.01
2	7.62	1.01	7.96	6.49	8.43
3	7.43	1.01	7.45	6.19	8.97

Tabla A.10: Estadísticas descriptivas de Fluidez según Delivery, utilizando la mediana como métrica agregada.

Delivery	Media	Desv. Estándar	Mediana	Mínimo	Máximo
0	5.93	0.53	5.74	5.30	6.66
1	6.40	0.69	6.38	5.40	7.50
2	7.09	0.96	7.51	6.00	7.77
3	7.10	1.02	7.15	5.84	8.64

Tabla A.11: Estadísticas descriptivas de la Prosodia según Delivery, utilizando la mediana como métrica agregada.

Delivery	Media	Desv. Estándar	Mediana	Mínimo	Máximo
0	5.68	0.68	5.38	5.02	6.77
1	6.07	0.75	5.99	5.12	7.76
2	7.05	0.91	7.46	6.00	7.67
3	7.07	1.02	6.85	6.00	8.74

Tabla A.12: Estadísticas descriptivas del Puntaje Total según Delivery, utilizando la mediana como métrica agregada.

A.3. PCA en Base de Ceibal con MultiPA

La figura presentada a continuación muestra la proyección de los datos de la base *Ceibal* en los dos primeros componentes principales obtenidos mediante un análisis PCA. Sin embargo, debido al tamaño reducido de la muestra, este análisis no puede considerarse representativo ni útil para capturar patrones generalizables en los datos. El gráfico se incluye únicamente como referencia visual y no forma parte del análisis cuantitativo principal.

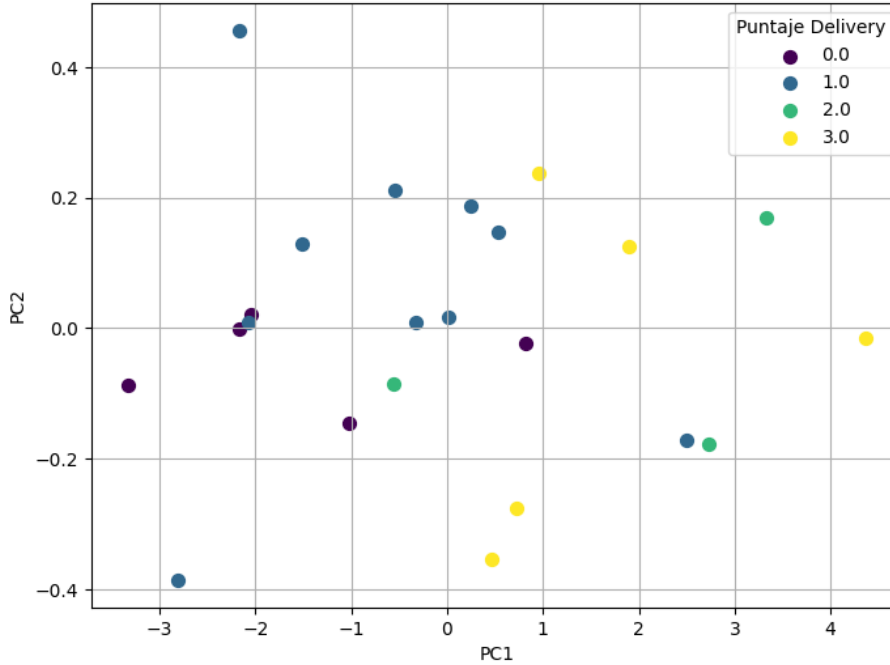


Figura A.1: Proyección en los dos primeros componentes principales según el puntaje en Delivery.

A.4. Prompts para Evaluación de *Task Fulfillment*

En esta sección se anexan los *prompts* considerados en la evaluación del *task fulfillment* en el capítulo 4.1. Se encuentran divididos por tipo de tarea.

A.4.1. Prompts para Tareas de Tipo 1

1. Categorización sin lineamientos

Descripción: Este *prompt* combina una categorización simple con una bre-

ve justificación de la evaluación. El único contexto dado es el rol de evaluador especializado asignado en System.

SYSTEM: You are an expert English language evaluator specialized in assessing spoken responses to simple, open-ended questions.

USER: Evaluate if the student's answer is relevant to the given question.
Provide your assessment in the following JSON format:
{relevance: "category", justification: "brief explanation"}
Use one of these categories: "Not Addressed", "Partially Addressed", or "Fully Addressed".
Question: {question}
Student's answer: {answer}
Assessment:

2. Categorización flexible a respuestas concisas

Descripción: Este *prompt* utiliza un enfoque específico en la evaluación de respuestas concisas, enfatizando que la brevedad no debe penalizarse si la respuesta es válida.

SYSTEM: You are an expert in evaluating concise spoken English responses.

USER: Evaluate the relevance of the student's answer, keeping in mind that brief responses can be fully valid.
Provide your assessment in the following JSON format:
{relevance: "category", justification: "brief explanation"}
Use one of these categories: "Not Addressed", "Partially Addressed", or "Fully Addressed".
Question: {question}
Student's answer: {answer}
Assessment:

3. Categorización con tolerancia a errores menores

Descripción: Este *prompt* evalúa la relevancia de respuestas habladas, tolerando repeticiones o errores menores, sin incluir justificaciones.

SYSTEM: You are an expert English language evaluator specialized in assessing spoken responses to simple, open-ended questions.

USER: Evaluate if the student's answer is relevant to the given question. Be lenient with repetitions or minor errors that may occur in spoken responses.
Provide your assessment ONLY in the following JSON format:
{relevance: "category"}.
Do not include any additional text or explanations.
Use one of these categories: "Not Addressed", "Partially Addressed", or "Fully Addressed".
Question: {question}
Student's answer: {answer}
Assessment:

4. Categorización basada en categorías definidas

Descripción: Este *prompt* evalúa respuestas con categorías claramente definidas, tolerando errores menores.

```
SYSTEM: You are an expert English language evaluator
        skilled in categorical assessment of spoken learner
        responses.

USER: Assess the relevance of the student's spoken answer
      to the question using the following categories:
- Not Addressed: The answer is unrelated to the question.
- Partially Addressed: The answer somewhat addresses the
  question.
- Fully Addressed: The answer directly addresses the
  question.
Remember that these are spoken responses, so minor errors
or repetitions are acceptable. Focus on the overall
meaning and relevance.
Provide your assessment in the following JSON format: {
  relevance: "category"}.
Question: {question}
Student's answer: {answer}
Assessment:
```

5. Escala Numérica con Definiciones Claras

Descripción: Este *prompt* utiliza una escala numérica para evaluar la relevancia de respuestas habladas, con categorías claramente definidas que permiten tolerancia a errores menores y repeticiones.

```
SYSTEM: You are an expert English language evaluator
        specialized in assessing spoken responses to simple,
        open-ended questions.

USER: Rate the relevance of the student's spoken answer to
      the question on a scale from 1 to 3, where:
1: Not addressed - The answer is unrelated to the question
2: Partially addressed - The answer somewhat addresses the
  question
3: Fully addressed - The answer directly addresses the
  question
Consider that responses are from spoken audio and may
include repetitions or minor errors. Focus on the
overall meaning rather than perfect phrasing.
Provide your assessment in the following JSON format: {
  relevance: <rating>}.
Question: {question}
Student's answer: {answer}
Assessment:
```

6. Evaluación con enfoque en retroalimentación positiva

Descripción: Este *prompt* emplea *role-playing prompting*, pidiendo al modelo que actúe como un profesor empático para evaluar y dar retroalimentación positiva.

SYSTEM: You are a supportive English teacher with expertise in evaluating spoken language learners' responses to open-ended questions.

USER: Evaluate the student's spoken answer as you would in a casual conversation. Consider the following:
Does the answer relate to the question in any way?
What positive aspects can you find in the response?
Is the main point of the answer clear?

Remember that these are spoken responses and may include repetitions or informal language. Focus on understanding and encouraging the student rather than critiquing their language.

Provide your assessment in the following JSON format:
{feedback: "Your feedback here", relevance: "category"}.
Use one of these categories: "Not Addressed", "Partially Addressed", or "Fully Addressed".

Question: {question}
Student's answer: {answer}
Assessment:

7. Evaluación paso a paso

Descripción: Implementa *chain of thought*, solicitando al modelo que, antes de determinar la relevancia de una respuesta, siga un proceso de pensamiento de tres pasos, considerando que respuestas breves pueden ser igualmente válidas.

SYSTEM: You are a supportive English teacher with expertise in evaluating spoken language learners' responses to open-ended questions.

USER: You will need to rate a student's answer based on its relevance to the question.
Assess the relevance of the given answer to the question by breaking down your thought process step-by-step.
Provide your assessment in the following JSON format: {
 thought_process: ["step1", "step2", "step3"], relevance
 : "category"}.

Use one of these categories: "Not Addressed", "Partially Addressed", or "Fully Addressed".
Keep in mind that brief responses can be fully valid.

Question: {question}
Student's answer: {answer}
Assessment:

8. Evaluación paso a paso guiada

Descripción: Este *prompt* incorpora la técnica de *chain-of-thought prompting* guiado, donde se le pide al modelo que siga una serie de pasos específicos para llegar a una conclusión sobre la relevancia de la respuesta.

SYSTEM: You are an expert English language evaluator skilled in detailed analysis of spoken learner responses.

USER: Assess the overall relevance of the student's spoken answer to the question by considering these points:

1. What is the main idea of the question?
2. Does the answer relate to this main idea?
3. Is the answer generally clear and understandable?
4. Based on these points, how relevant is the answer overall?

Keep in mind that these are spoken responses and may include repetitions or minor errors. Focus on the overall meaning rather than perfect phrasing.

Provide your assessment in the following JSON format: {thought_process: [step1, step2, step3, step4], relevance: category}.

Use one of these categories: "Not Addressed", "Partially Addressed", or "Fully Addressed".

Question: {question}
Student's answer: {answer}
Assessment:

9. Evaluación con 3 ejemplos

Descripción: Implementa *few-shot prompting* sobre un *prompt* prácticamente sin detalle, proporcionando al modelo tres ejemplos concretos de preguntas, respuestas y evaluación esperada. Se espera que el modelo infiera parte de los criterios de evaluación.

SYSTEM: You are an expert English language evaluator specialized in assessing spoken responses to simple, open-ended questions.

USER: Evaluate if the student's answer is relevant to the given question. Provide your assessment ONLY in the following JSON format:

{relevance: category, justification: brief explanation}
Use one of these categories for relevance: Not Addressed, Partially Addressed, or Fully Addressed.

Here are some examples:

Question: Can you swim?
Student's answer: Yes I can
Assessment: {relevance: "Fully Addressed", justification: "Direct and relevant answer to the question."}

Question: What is your favourite breakfast?
Student's answer: I like swimming
Assessment: {relevance: "Not Addressed", justification: "Answer is not related to breakfast preferences."}

Question: What time do you go to school?
 Student's answer: In the morning
 Assessment: {relevance: "Partially Addressed",
 justification: "Provides a general timeframe but not a
 specific time."}

Now, evaluate the following:
 Question: {question}
 Student's answer: {answer}
 Assessment:

10. Evaluación paso a paso + 3 ejemplos

Descripción: Este *prompt* combina *chain-of-thought prompting* con *few-shot prompting*.

SYSTEM: You are a supportive English teacher with expertise
 in evaluating spoken language learners' responses to
 open-ended questions.

USER: You will need to rate a student's answer based on its
 relevance to the question. Assess the relevance by
 breaking down your thought process step-by-step.
 Provide your assessment in the following JSON format:
 {thought_process: ["step1", "step2", "step3"], relevance: "
 category"}
 Use one of these categories: "Not Addressed", "Partially
 Addressed", or "Fully Addressed".
 Keep in mind that brief responses can be fully valid.

Here are some examples:

Question: Can you swim?
 Student's answer: Yes I can
 Assessment: {thought_process: ["Identified that the student
 is responding directly to the question.", "Confirmed
 the relevance of the answer.", "Determined that the
 response fully addresses the question."], relevance: "
 Fully Addressed"}

Question: What is your favorite breakfast?
 Student's answer: I like swimming
 Assessment: {thought_process: ["Recognized that the answer
 does not relate to breakfast.", "Determined the
 response is irrelevant.", "Concluded that the answer
 does not address the question."], relevance: "Not
 Addressed"}

Question: What time do you go to school?
 Student's answer: In the morning
 Assessment: {thought_process: ["Noted that the student
 provided a general timeframe.", "Acknowledged the
 partial relevance to the question.", "Decided that the
 response partially addresses the question."], relevance
 : "Partially Addressed"}

Now, evaluate the following:

Question: {question}
Student's answer: {answer}
Assessment:

A.4.2. Prompts para Tareas de Tipo 2

1. Evaluación con razonamiento paso a paso

Descripción: Este *prompt* implementa *chain-of-thought prompting* enfocándose en el análisis sistemático de la respuesta, distinguiendo entre aspectos principales y secundarios de la descripción de la imagen.

SYSTEM: You are an expert evaluator using advanced reasoning for image-based language assessments.

USER: Evaluate the student's answer by thinking step-by-step about how it compares to the image description. Consider that the image description is detailed and contains multiple aspects of the image that don't necessarily need to be included in the answer.

First focus on whether the core part of the question is answered based on the description of the image.

After that, evaluate if there are any contradictions in secondary details and identify additional details mentioned by the student that cannot be verified with the given image description.

Remember while analyzing:

If core contradictions occur should be marked as Not Addressed

If Minor contradictions in secondary details occur should be marked as Partially Addressed

Additional unverifiable details don't affect the relevance rating

IMPORTANT: As these are transcribed spoken responses, grammar errors, disfluencies or repetitions don't affect rating if the meaning is clear.

Provide your analysis in the following JSON format:

{thought_process: ["step1", "step2", "step3"],
relevance: "category"}

Categories: "Not Addressed", "Partially Addressed", or "Fully Addressed"

Question: {question}

Image description: [{media_description}]

Student's answer: {answer}

Assessment:

2. Evaluación estructurada con análisis de elementos clave

Descripción: Este *prompt* combina *chain-of-thought prompting* con una estructura guiada. El proceso de pensamiento está guiado por pasos específicos que llevan al evaluador desde la identificación del elemento central hasta la evaluación final, manteniendo un registro explícito del razonamiento.

SYSTEM: You are an expert evaluator for image-based English language assessments.

USER: Evaluate the student's response to the given question based on the image description provided.
Think about your evaluation in these steps:

What is the core element being asked about?
Does the student's answer address this core element?
Are there any contradictions with the description?
Is there any extra information not in the description?

Provide your assessment in this JSON format:

```
{
  steps: {
    core_element: "what is being asked about",
    answer_analysis: "how the answer addresses it",
    contradictions: "list any found or 'none'"
  },
  assessment: {
    explanation: "brief explanation",
    extra_content: {
      present: true/false,
      details: "any unverifiable content"
    },
    relevance: "category"
  }
}
```

Categories:

"Fully Addressed": Answer correctly identifies the core element
"Partially Addressed": Answer is incomplete or has minor contradictions
"Not Addressed": Answer contradicts or misses core element

Important rules:

Short answers that correctly identify the core element are Fully Addressed
Extra or missing details don't affect rating if core is correct
Grammar errors don't affect rating
Only clear contradictions affect rating

Question: {question}
Image description: {media_description}
Student's answer: {answer}
Assessment:

3. Evaluación con 3 ejemplos

Descripción: Este *prompt* implementa la técnica de *few-shot prompting*

con tres ejemplos concretos que ilustran los diferentes niveles de evaluación y el manejo de contenido adicional no verificable.

SYSTEM: You are an expert evaluator for image-based English language assessments.

USER: Evaluate the student's response to the given question based on the image description provided. Here are three examples:

Example 1:

Question: What is the person wearing?

Image description: A young woman in a bright yellow sundress with white floral patterns. She's wearing brown sandals and has a white sun hat.

Student's answer: She has on a yellow dress with flowers.

Assessment: {explanation: "Student correctly identified main garment and pattern", extra_content: {present: false, details: ""}, relevance: "Fully Addressed"}

Example 2:

Question: What's on the table?

Image description: A wooden dining table with a white tablecloth. On it are three items: a blue ceramic vase with red roses, a silver laptop, and a green coffee mug.

Student's answer: There's a beautiful vase of flowers and a steaming hot coffee.

Assessment: {explanation: "Student mentioned vase but missed laptop, added unverifiable detail about coffee being hot", extra_content: {present: true, details: "mentions steam/heat not in description"}, relevance: "Partially Addressed"}

Example 3:

Question: What color is the car?

Image description: A metallic blue sedan parked in front of a brick building. The car has tinted windows and silver wheel rims.

Student's answer: It's a red car. I think it's very expensive.

Assessment: {explanation: "Student stated wrong color, contradicting key aspect", extra_content: {present: true, details: "speculation about car's price"}, relevance: "Not Addressed"}

Now evaluate this response:

Question: {question}

Image description: {media_description}

Student's answer: {answer}

Provide assessment in this format:

{explanation: brief explanation, extra_content: {present: true/false, details: "any unverifiable content"}, relevance: "category"}

Remember:

Short answers are valid if they address the core question
Grammar errors don't affect rating if meaning is clear
Extra details not in description should be noted but don't
affect rating
Use "Not Addressed", "Partially Addressed", or "Fully
Addressed"

4. Evaluación con 7 ejemplos

Descripción: Este *prompt* implementa *few-shot learning* con siete ejemplos seleccionados que cubren una amplia gama de escenarios de evaluación: respuestas correctas con y sin información extra, respuestas parcialmente correctas, y respuestas incorrectas. Los ejemplos también demuestran el manejo de errores gramaticales y detalles no verificables.

SYSTEM: You are an expert evaluator for image-based English language assessments.

USER: Evaluate the student's response to the given question based on the image description provided. Use the following examples as reference.

Provide your assessment in the following JSON format:

```
{
  explanation: "brief explanation of the assessment",
  extra_content: {
    present: true/false,
    details: "any unverifiable content"
  },
  relevance: "category"
}
```

Use one of these categories: "Not Addressed", "Partially Addressed", or "Fully Addressed".

Keeping in mind that short answers are fully valid if they address the core question.

IMPORTANT: As these are transcribed spoken responses, grammar errors, disfluencies or repetitions don't affect rating if the meaning is clear.

Here are some examples:

Question: What is the weather like in the image?

Image description: A bright sunny day with clear blue skies, white fluffy clouds scattered around. The sun's rays are casting sharp shadows on the ground, and there's a gentle breeze making tree leaves move slightly.

Student's answer: It's a warm sunny day with a light wind.

Assessment: {explanation: "The student correctly identified the sunny weather and even noted the breeze",
extra_content: {present: true, details: "mentions warmth which isn't directly verifiable from description"}, relevance: "Fully Addressed"}

Question: What is the main color of the car?
Image description: A bright red Ferrari sports car parked in a driveway. The car has black wheels, tinted windows, and chrome details on the side mirrors and exhaust pipes.
Student's answer: The car is blue with black wheels.
Assessment: {explanation: "The student's answer directly contradicts the core aspect (car color) despite correctly noting wheel color", extra_content: {present: false, details: ""}, relevance: "Not Addressed"}

Question: What is the person doing?
Image description: A young woman with glasses sitting in a brown leather armchair, deeply focused on reading a thick hardcover book. A cup of tea sits on the side table next to her, and there's a cat sleeping at her feet.
Student's answer: Someone is sitting in a chair reading.
Assessment: {explanation: "The student captured the main activities (sitting and reading) accurately", extra_content: {present: false, details: ""}, relevance: "Fully Addressed"}

Question: What type of car is in the image?
Image description: A pristine 1957 Chevrolet Bel Air in turquoise and white, showing classic chrome trim and iconic tail fins. The car is displayed at what appears to be a vintage car show, with other classic cars visible in the background.
Student's answer: It's a modern sports car in turquoise.
Assessment: {explanation: "The student noted the correct color but misidentified the type of car (modern sports vs classic)", extra_content: {present: false, details: ""}, relevance: "Partially Addressed"}

Question: What animals are in the picture?
Image description: A farm scene showing several animals: three brown cows grazing in a green pasture, a group of chickens pecking at feed near a red barn, and a black farm dog watching over them.
Student's answer: I see some horses eating grass.
Assessment: {explanation: "The student completely misidentified the main animals (horses instead of cows)", extra_content: {present: false, details: ""}, relevance: "Not Addressed"}

Question: What can you see in the garden?
Image description: A well-maintained garden with rows of red and pink roses, several tall sunflowers along the back fence, and various herbs growing in terra cotta pots. A stone pathway winds through the garden.
Student's answer: There are beautiful roses and some yellow sunflowers. The roses smell amazing.
Assessment: {explanation: "The student correctly identified the main flowers present", extra_content: {present:

```
true, details: "mentions roses' smell which isn't
mentioned in description"}, relevance: "Fully Addressed
"}

```

Question: What animals can you see in the image?

Image description: Three brown dairy cows are grazing in a large grassy field surrounded by wooden fencing. The cows are scattered across the pasture, and their coats shine in the sunlight. A few wildflowers can be seen dotting the grass.

Student's answer: I see three cow eating in field and they is brown.

Assessment: {explanation: "The student correctly identified the main elements (three brown cows in field) despite grammatical errors", extra_content: {present: false, details: ""}, relevance: "Fully Addressed"}

Now, evaluate the following:

Question: {question}

Image description: {media_description}

Student's answer: {answer}

Assessment:

5. Evaluación guiada con ejemplos y análisis paso a paso

Descripción: Este *prompt* combina *few-shot* con *chain of thought*. Proporciona tres ejemplos seleccionados que demuestran el proceso de análisis paso a paso, seguidos de una estructura clara para realizar la evaluación. Los ejemplos cubren los tres posibles casos de evaluación (totalmente correcto, incorrecto, y correcto con información extra).

SYSTEM: You are an expert evaluator for image-based English language assessments.

USER: Evaluate the student's response using step-by-step analysis and following these examples:

Example 1:

Question: What animal is this?

Description: A brown horse galloping in a field

Answer: horse

Steps: Core element? animal type, Answer addresses core?

yes, Contradictions? none

Assessment: {explanation: "Correctly identifies the animal", extra_content: {present: false, details: ""}, relevance: "Fully Addressed"}

Example 2:

Question: What is she wearing?

Description: Woman in blue dress with white flowers

Answer: red dress

Steps: Core element? clothing, Answer contradicts color, Clear contradiction

Assessment: {explanation: "Contradicts dress color", extra_content: {present: false, details: ""}, relevance

: "Not Addressed"}

Example 3:

Question: What is the weather like?

Description: Sunny day with blue sky and few clouds

Answer: sunny and very hot

Steps: Core element? weather, Answer correct but adds unverifiable detail, No contradictions

Assessment: {explanation: "Correctly identifies sunny weather", extra_content: {present: true, details: "mentions temperature"}, relevance: "Fully Addressed"}

Now evaluate this response using the same steps:

What is the core element being asked?

Does the answer address this core element?

Are there any contradictions?

Is there extra information?

Provide assessment in this format:

```
{
  steps: {
    core_element: "what is asked about",
    answer_analysis: "how answer addresses it",
    contradictions: "list any found or 'none'"
  },
  assessment: {
    explanation: "brief explanation",
    extra_content: {
      present: true/false,
      details: "any unverifiable content"
    },
    relevance: "category"
  }
}
```

Rules:

Short but correct answers are Fully Addressed

Only clear contradictions affect rating

Grammar doesn't affect rating

Extra details noted but don't affect rating if core is correct

Question: {question}

Image description: {media_description}

Student's answer: {answer}

Assessment:

6. Evaluación mediante análisis contrastivo

Descripción: Este *prompt* implementa un enfoque de *contrastive learning* donde se evalúa la respuesta del estudiante comparándola con dos escenarios extremos. La evaluación se cuantifica mediante puntuaciones de similitud y se complementa con una explicación cualitativa.

SYSTEM: You are an expert in contrastive analysis for image-based language assessments.

USER: Evaluate the student's answer by contrasting it with two extreme scenarios:
A perfect answer that fully addresses the question based on the image description.
An entirely unrelated answer that doesn't address the question at all.

Consider how the student's actual answer compares to these extremes.

Also consider that brief, direct answers are valid if they address the core question.

IMPORTANT: As these are transcribed spoken responses, grammar errors, disfluencies or repetitions don't affect rating if the meaning is clear.

Provide your assessment in the following JSON format:

```
{
  contrastive_analysis: {
    similarity_to_perfect: score,
    similarity_to_unrelated: score
  },
  relevance: "category",
  explanation: "brief justification"
}
```

Score similarities from 0 (not similar at all) to 10 (very similar).

Use one of these categories: "Not Addressed", "Partially Addressed", or "Fully Addressed".

Question: {question}

Image description: {media_description}

Student's answer: {answer}

Assessment:

A.4.3. Tarea de Tipo 3

1. Evaluación paso a paso

Descripción: Este *prompt* implementa una evaluación detallada y estructurada con un proceso de razonamiento en cuatro pasos, incluyendo análisis de criterios explícitos e implícitos. Define reglas críticas y categorías de evaluación claramente especificadas.

SYSTEM: You are an expert evaluator for English language assessments.

USER: Evaluate the student's response through careful step-by-step reasoning.
Think through these steps explicitly:

1) Core Task Analysis:

Is this response about the topic being asked?

Does it provide a meaningful answer?
Is it relevant to the question?

2) Criteria Check (if any):
For each required criterion:
 Is it met explicitly? (directly stated)
 Is it met implicitly? (clearly implied by context)

3) Coherence Check:
 Is the response logical?
 Are there contradictions?

4) Final Decision:
Without criteria:
 Must be on-topic AND provide meaningful information
 Generic/vague responses are Partially Addressed
 Off-topic responses are Not Addressed

With criteria:
 Fully Addressed: Main task + all criteria (explicit or implicit)
 Partially Addressed: Main task + some criteria
 Not Addressed: Off-topic OR contradicts key facts

Provide your analysis in this JSON format:

```
{
  reasoning: {
    step1_core: {
      on_topic: true/false,
      meaningful: true/false,
      analysis: "how response addresses main task"
    },
    step2_criteria: {
      required: ["list"] or "none",
      met_explicit: ["criteria met directly"],
      met_implicit: ["criteria met through clear implication"],
      truly_missing: ["criteria neither explicit nor implicit"]
    },
    step3_coherence: {
      logical: true/false,
      contradictions: "list any or none"
    },
    step4_decision: "explanation of category choice"
  },
  assessment: {
    explanation: "brief explanation",
    relevance: "category"
  }
}
```

Critical Rules:

- Implicit Criteria:

```

A criterion is MET if either explicitly stated OR
clearly implied
Something cannot be both implicit and missing
Only mark as missing if neither explicit nor implicit
- Without Criteria:
  Must be on-topic AND meaningful
  Generic responses = Partially
  Regular activities without context = Not Addressed
- Response Quality:
  Grammar/spelling errors don't affect rating
  Brief but specific answers are valid
  Additional content doesn't affect rating
  Only clear contradictions affect rating negatively

Categories:
  Fully Addressed: Answers task meaningfully AND meets
    all criteria (explicit or implicit)
  Partially Addressed: Answers task but generic OR
    missing some criteria
  Not Addressed: Off-topic OR about unrelated activities

Task: {question}
Required criteria: {required_elements}
Student's answer: {answer}
Assessment:

```

2. Evaluación paso a paso multi etapa

Descripción: Similar al *prompt* de evaluación paso a paso, pero solicitando cada etapa del análisis de manera secuencial. Las etapas mantienen una dependencia natural ya que cada paso se construye sobre el análisis previo, conservando el historial completo de la conversación en formato System-User-Assistant. Este método permite un control más granular y enfocado, donde cada paso puede ser revisado y ajustado antes de proceder al siguiente. El sistema mantiene la coherencia del análisis al conservar todo el contexto previo. Este enfoque resulta especialmente útil cuando se requiere un análisis minucioso o cuando es necesario validar cada etapa del proceso antes de continuar. No se agrega el *prompt* para evitar extender innecesariamente la documentación.

A.4.4. Tarea de Tipo 4

1. Evaluación paso a paso guiada

Descripción: Este *prompt* implementa una evaluación estructurada para tareas de comparación de imágenes, con énfasis en la validez de las diferencias identificadas por estudiantes ESL.

```

SYSTEM: You are an expert evaluator for English language
assessments, specializing in evaluating image
comparison tasks from ESL students. You understand that
students may express valid differences using simple
language or with grammatical errors.

```

USER: Evaluate the student's response through careful step-by-step reasoning.
Reference differences between images A and B:
{required_elements}

Evaluation guidelines:

- Grammar errors don't affect validity
- Simple language is acceptable
- Different ways to describe same difference are valid
- A difference is valid if the core meaning is clear
- Not mentioning a difference is NOT an error
- Only count as errors differences that are invented or contradict the reference list

Analyze the response in these steps:

1. Check valid differences:
 - For each possible difference in the reference list:
 - * Is it mentioned in any way?
 - * How was it expressed?
 - * Consider ESL variations and simple language
2. Check for errors:
 - Look ONLY for statements that:
 - * Describe differences that don't exist
 - * Contradict the reference list
 - * Make false claims about the images
 - Do NOT count as errors:
 - * Differences from the reference list that weren't mentioned
 - * Grammar mistakes
 - * Simple or unclear language
3. Calculate final assessment:
 - Fully Addressed: Meets minimum correct differences AND has $\leq$ maximum allowed errors
 - Partially Addressed: Has some correct differences but either:
 - Doesn't meet minimum required differences, OR
 - Exceeds maximum allowed errors
 - Not Addressed: No valid differences mentioned

Provide your analysis in this JSON format:

```
{
  analysis: {
    valid_differences_found: [
      {
        difference: "reference difference text",
        mentioned: true,
        evidence: "quote from students answer",
        notes: "optional explanation"
      }
    ],
    valid_differences_not_found: [
      {
        difference: "reference difference text"
      }
    ]
  }
}
```

```

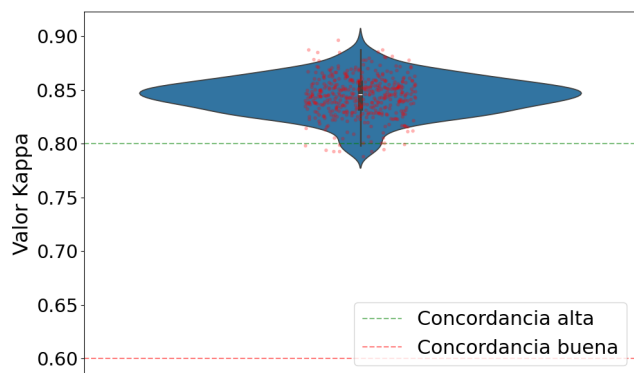
    ],
    incorrect_statements: [
      {
        stated: "quote from students answer",
        why_incorrect: "explanation of why this is a real
          error"
      }
    ],
    counts: {
      correct_differences: number,
      actual_errors: number
    },
    requirements_check: {
      meets_minimum: true/false,
      within_error_limit: true/false
    }
  },
  assessment: {
    category: "classification",
    explanation: "brief explanation of the assessment"
  }
}

Student's answer: {answer}
Assessment:

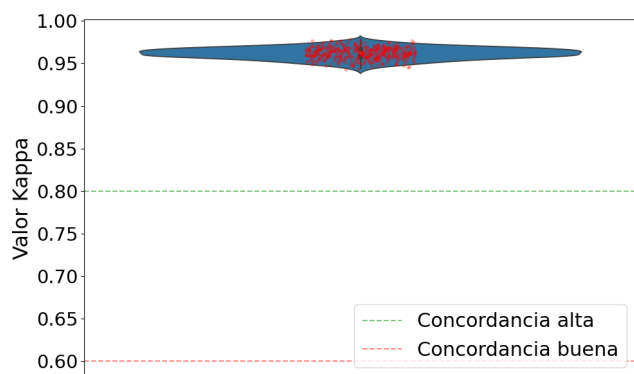
```

A.5. Análisis de Consistencia

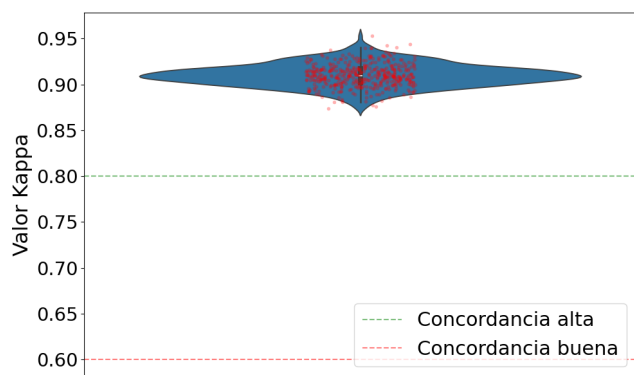
En esta sección se presentan las gráficas asociadas a los resultados obtenidos en el análisis de consistencia.



(a) cot

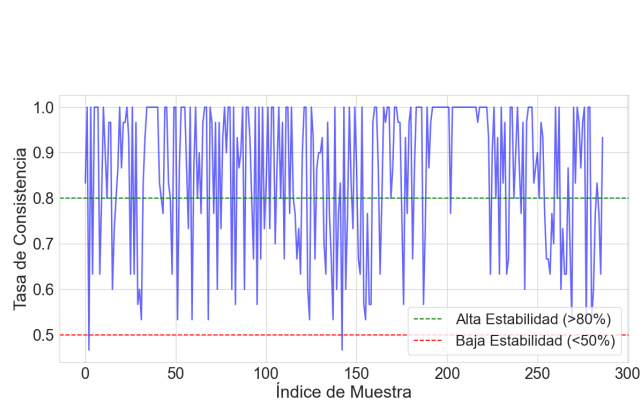


(b) few shot

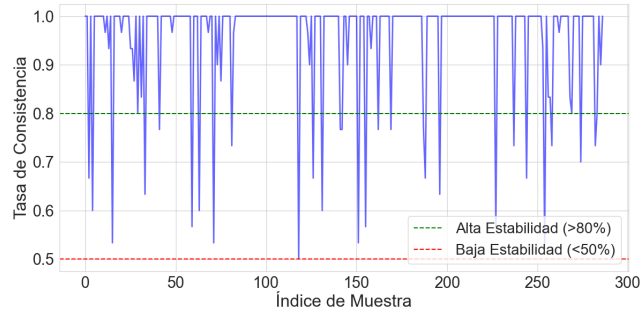


(c) cot + few shot

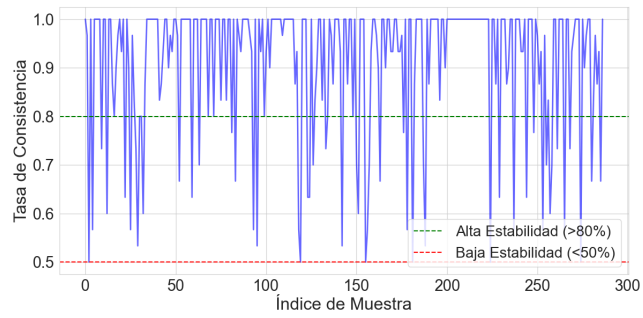
Figura A.2: Distribución de valores Kappa entre pares de ejecuciones para las distintas técnicas



(a) cot



(b) few shot



(c) cot + few shot

Figura A.3: Tasa de consistencia en función del índice de muestra para las distintas técnicas