EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Pre-procesamiento de *Prompt's* para Modelos de Lenguaje Grandes

presentada por
Ing. Carlos Javier Martínez Joffroy

como requisito para la obtención del grado de
Maestro en Ciencias de la Computación

Director de tesis
Dr. Nimrod González Franco

Co-directora de tesis
Dra. Andrea Magadán Salazar

Comité tutorial
Dr. Raúl Pinto Elías
Dr. Noé Alejandro Castro Sánchez

Cuernavaca, Morelos, México. Diciembre del 2025.

cenidet®
Centro Nacional de Investigación y Desarrollo Tecnológico

Centro Nacional de Investigación y Desarrollo tecnológico
Departamento de Ciencias Computacionales

Cuernavaca, Mor., 11/agosto/2025
OFICIO No. DCC/219/2025
Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial de **Carlos Javier Martínez Joffroy** con número de control **M23CE068**, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Pre-procesamiento de Prompt's para Modelos de Lenguaje Grandes"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

ATENTAMENTE

Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"

Dr. Nimrod González Franco
Director de tesis

Dra. Andrea Magadán Salazar
Codirectora de tesis

Dr. Raúl Pinto Elías
Revisor 1

Dr. Noé Alejandro Castro Sánchez
Revisor 2

C.c.p. Depto. Servicios Escolares.
Expediente / Estudiante

Centro Nacional de Investigación y Desarrollo tecnológico
Subdirección Académica

Cuernavaca Mor, 01/diciembre/2025
Oficio No. SAC/276/2025
Asunto: Autorización de impresión de tesis

CARLOS JAVIER MARTÍNEZ JOFFROY
CANDIDATO AL GRADO DE MAESTRO
EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Pre-procesamiento de Prompt´s para Modelos de Lenguaje Grandes"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz

2025 Año de La Mujer Indígena

DEDICATORIA

A mi querida madre y hermana, por su apoyo incondicional, paciencia y amor durante todo mi desarrollo educativo a lo largo de estos dos años de maestría.

Gracias por ser mi fortaleza y guía en cada paso de este camino.

AGRADECIMIENTOS

Agradezco a la Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) por el apoyo económico brindado durante mis estudios de maestría.

Al Tecnológico Nacional de México / Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET) por brindar las instalaciones y permitirme realizar los estudios de maestría.

A mi director de tesis, el Dr. Nimrod González Franco, a la co-directora de tesis, Dra. Andrea Magadán Salazar y a los miembros del comité tutorial, el Dr. Raúl Pinto Elías y el Dr. Noé Alejandro Castro Sánchez, por su guía, conocimientos y valiosas recomendaciones, que fueron esenciales para el desarrollo y mejora de este trabajo.

RESUMEN

Actualmente, la calidad de las interacciones entre los usuarios y los Modelos de Lenguaje es fundamental para alcanzar resultados óptimos en diversas aplicaciones. La experiencia demuestra que dicha calidad depende en gran medida de los *Prompts* empleados. Sin embargo, la literatura existente carece de estudios que evalúen de forma objetiva y cuantitativa la relación entre los *Prompts* y las respuestas generadas por un Modelo Grande de Lenguaje. Esto se debe, en parte, a la ausencia de métricas adecuadas para valorar la calidad de los textos de entrada, lo cual limita la evaluación precisa de los *Prompts*. En este contexto, el presente documento introduce PromptIQ, una métrica diseñada para medir la calidad de los *Prompts* y proporcionar un marco de referencia que permita la investigación sobre la relación entre las entradas y salidas de un Modelo Grande de Lenguaje.

Palabras clave: promptiq, modelos grandes de lenguaje, calidad de *prompts*.

ABSTRACT

Currently, the quality of interactions between users and Language Models is crucial for achieving optimal results across various applications. Evidence shows that such quality largely depends on the *Prompts* employed. However, the existing literature lacks studies that objectively and quantitatively evaluate the relationship between *Prompts* and the responses generated by a Large Language Model. This is partly due to the absence of appropriate metrics to assess the quality of input texts, which limits the precise evaluation of *Prompts*. In this context, the present document introduces PromptIQ, a metric designed to measure the quality of *Prompts* and provide a reference framework for investigating the relationship between the inputs and outputs of a Large Language Model.

Keywords: promptiq, large language models, *prompt* quality.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ACRÓNIMOS

Acrónimo	Significado en Inglés	Significado en Español
ACC	*Accuracy*	Exactitud, proporción de predicciones correctas sobre el total
ANN	*Artificial Neural Network*	Red Neuronal Artificial
APE	*Automatic Prompt Engineering*	Ingeniería de *Prompts* automática
ATTSiaBiLSTM	*Attention-based Siamese Bidirectional Long Short-Term Memory*	Memoria a Corto y Largo Plazo Bidireccional Siamés con Mecanismo de Atención
AUC	*Area Under the Curve*	Área bajo la curva
BART	*Bidirectional and Auto-Regressive Transformers*	Transformadores Bidireccionales y Auto-regresivos
BERT	*Bidirectional Encoder Representations from Transformers*	Representaciones Codificadas Bidireccionales a partir de Transformadores
BERTScore	*Bidirectional Encoder Representations from Transformers*	Puntaje BERT
BLEU	*Bilingual Evaluation Understudy*	Evaluador Bilingüe Automático
BoW	*Bag of Words*	Bolsa de Palabras
BP	*Brevity Penalty*	Penalización por brevedad
BPE	*Byte Pair Encoding*	Codificación por Pares de Bytes

ACRÓNIMOS

Acrónimo	Significado en Inglés	Significado en Español
CENIDET		Centro Nacional de Investigación y Desarrollo Tecnológico
CIDEr	*Consensus-based Image Description Evaluation*	Evaluación de Descripción de Imagen Basada en Consenso
CNN	*Convolutional Neural Network*	Red Neuronal Convolucional
Conv	*Conversational Model*	Modelo Conversacional
CoOP	*Context Optimization for Prompting*	Optimización de Contexto para *Prompts*
CoT	*Chain-of-Thought*	Cadena de razonamiento
CPM-2	*Chinese Pre-trained Models*	Modelos Pre-entrenados chinos
CTRL	*Conditional Transformer Language Model*	Modelo de Transformer Condicional
DPR	*Dense Passage Retrieval*	Recuperación Densa de Pasajes
DT	*Decision Tree*	Árbol de Decisión
E2E	*End-to-End*	Extremo a Extremo
ELMo	*Embeddings from Language Models*	Incrustaciones desde un Modelo de Lenguaje

ACRÓNIMOS

Acrónimo	Significado en Inglés	Significado en Español
ERNIE	*Enhanced Representation Through Knowledge Integration*	Representación Mejorada Mediante Integración de Conocimiento
FN	*False Negatives*	Falsos negativos
FP	*False Positives*	Falsos positivos
FT	*Fine-tuning*	Ajuste fino
GLM	*General Language Model*	Modelo de Lenguaje General
GPEI	*Generative Prompt Engineering Iterative*	Ingeniería Iterativa de *Prompts* Generativos
GPT	*Generative Pre-trained Transformer*	Transformador Generativo Preentrenado
IPL	*Instance-aware Prompt Learning*	Aprendizaje de *Prompts* Sensibles a la Instancia
KDE	*Kernel Density Estimation*	Estimación de Densidad con Núcleo
L2R	*Left-to-Right*	De Izquierda a Derecha
LDA	*Latent Dirichlet Allocation*	Asignación Latente de Dirichlet
LLM	*Large Language Model*	Modelo Grande de Lenguaje

ACRÓNIMOS

Acrónimo	Significado en Inglés	Significado en Español
LM	*Language Model*	Modelo de Lenguaje
LR	*Logistic Regression*	Regresión Logística
LSTM	*Long Short-Term Memory*	Memoria a Largo y Corto Plazo
MALLET	*Machine Learning for Language Toolkit*	Herramienta para Aprendizaje Automático de Lenguaje
MCC	*Matthews Correlation Coefficient*	Coeficiente de correlación de Matthews
METEOR	*Metric for Evaluation of Translation with Explicit ORdering*	Métrica para Evaluación de Traducción con Orden Explícito
Mine+Man	*Combination of automatic Prompt Mining and manual paraphrasing to generate effective variants*	Combinación de minería automática y parafraseo manual de *Prompts* para generar variantes efectivas
MSE	*Mean Squared Error*	Error cuadrático medio, medida de diferencia entre valores predichos y reales
NB	*Naive Bayes*	Naive Bayes
NER	*Named Entity Recognition*	Reconocimiento de Entidades Nombradas
NIR	*Neural Information Retrieval*	Recuperación de Información Neuronal
NIST	*National Institute of Standards and Technology*	Instituto Nacional de Estándares y Tecnología

ACRÓNIMOS

Acrónimo	Significado en Inglés	Significado en Español
NLI	*Natural Language Inference*	Inferencia de Lenguaje Natural
NLP	*Natural Language Processing*	Procesamiento de Lenguaje Natural
NMF	*Non-negative Matrix Factorization*	Factorización de Matrices No Negativas
NMT	*Neural Machine Translation*	Traducción Automática Neuronal
OCTIS	*Optimizing Comparisons of Topic models Is Simple*	Optimización de Comparación de Modelos de Temas de Forma Simple
ONE-HOT	*One-hot encoding*	Codificación binaria de categorías en vectores dispersos
OPTI	*Optimal Prompt selection based on combination or evaluation*	Selección óptima de *Prompts* basada en combinación o evaluación
PCA	*Principal Component Analysis*	Análisis de Componentes Principales
PEGASUS	*Pre-training with Extracted Gap-sentences for Abstractive Summarization*	Pre-entrenamiento con Oraciones Faltantes para Resumen automático
PEPLER-D	*Personalized Prompt Learning for Explainable Recommendation – Discrete Prompts*	Aprendizaje de *Prompts* Personalizados para Recomendación Explicable – *Prompts* Discretos
PI	*Paraphrase Identification*	Paráfrasis
PLM	*Pretrained Language Models*	Modelos de Lenguaje Preentrenados

ACRÓNIMOS

Acrónimo	Significado en Inglés	Significado en Español
POS	*Part-of-Speech*	Partes del Discurso
PPL	*Perplexity*	Perplejidad
PT	*Prompt Tuning*	Ajuste de *Prompt*
Q2Q	*Question-to-Question Similarity*	Similitud de Pregunta a Pregunta
RAKE	*Rapid Automatic Keyword Extraction*	Extracción automática de palabras clave mediante análisis de frecuencia y co-ocurrencia
RAVEN	*Ranked Agreement of Variability in Example Novelties*	Acuerdo Ordenado de la Variabilidad en Novedades de Ejemplos
ReAct	*Reason Act*	Razonamiento y Acción
ROUGE	*Recall-Oriented Understudy for Gisting Evaluation*	Evaluación de Resumen Orientada al Recuerdo
SECIHTI		Secretaría de Ciencia, Humanidades, Tecnología e Innovación
SiaBiLSTM	*Siamese Bidirectional Long Short-Term Memory*	Memoria a Corto y Largo Plazo Bidireccional Siamés
SiaLSTM	*Siamese Long Short-Term Memory*	Memoria a Corto y Largo Plazo Siamés
SMT	*Statistical Machine Translation*	Traducción Automática Estadística

ACRÓNIMOS

Acrónimo	Significado en Inglés	Significado en Español
SQAS	*Short Question Answering Systems, designed to answer brief questions accurately using language models*	Sistemas de preguntas cortas, diseñados para responder preguntas breves con precisión usando modelos de lenguaje
STSS	*Short-Text Semantic Similarity*	Similitud Semántica de Textos Cortos
SVM	*Support Vector Machine*	Máquina de Vectores de Soporte
SWT	Shapiro-Wilk *Test*	Prueba de Shapiro-Wilk
T5	*Text-to-Text Transfer Transformer*	Transformador de Texto a Texto
TER	*Translation Edit Rate*	Tasa de Edición de Traducción
TF-IDF	*Term Frequency – Inverse Document Frequency*	Frecuencia de Término – Frecuencia Inversa de Documento
Top1	*Top-1 selection according to main metric*	Mejor opción seleccionada según la métrica principal
TP	*True Positives*	Verdaderos positivos
Trans	*Transformer*	Transformador
VAETM	*Variational Autoencoder Topic Model*	Modelo de Temas con Autoencoder Variacional
VSM	*Vector Space Model*	Modelo de espacio vectorial para representación de documentos

ACRÓNIMOS

Acrónimo	Significado en Inglés	Significado en Español
WST	Wilcoxon *Signed-Rank Test*	Prueba de rangos con signo de Wilcoxon

Capítulo 1

Introducción

Los Modelos Grandes de Lenguaje (del inglés *Large Language Models* o LLMs) se han consolidado como herramientas fundamentales en aplicaciones de procesamiento de lenguaje natural (del inglés *Natural Language Processing*, NLP), incluyendo generación de texto, análisis semántico, traducción automática y extracción de información. Sin embargo, se plantea que la forma en que se redactan los *Prompts* podría estar relacionada con las características de las respuestas obtenidas, aunque esta relación aún requiere mayor estudio. Por esta razón, surge la necesidad de proponer estrategias que permitan evaluar y mejorar la calidad de los *Prompts* antes de enviarlos a un LLM, favoreciendo así interacciones más efectivas entre humanos y máquinas.

1.1 Planteamiento del Problema

En la actualidad, la eficacia de un LLM puede verse influida por la calidad de los *Prompts* que recibe. La forma en que se redactan las entradas es un aspecto relevante a considerar para optimizar la interacción con los LLMs. Los usuarios no expertos carecen, generalmente, de las habilidades necesarias para redactar *Prompts* efectivos, lo que plantea la necesidad de estudiar estrategias para mejorar su calidad. Otro aspecto relevante es la limitada disponibilidad de métricas que permitan evaluar de manera sistemática la calidad de los *Prompts*, lo que dificulta cuantificar cómo las entradas influyen en el comportamiento de los LLM.

1.2 Delimitación del Problema

Esta tesis se centra en evaluar y mejorar la calidad de los *Prompts* redactados por usuarios no expertos en español culto, como etapa de pre-procesamiento antes de ser utilizados como entrada en un LLM.

La investigación se limita a *Prompts* redactados por usuarios no expertos en español, excluyendo entradas generadas automáticamente o en otros idiomas. Se enfoca en identificar ca-

racterísticas de los *Prompts* que puedan optimizarse para favorecer interacciones más efectivas entre usuarios y LLMs. Asimismo, se busca establecer criterios que permitan evaluar de manera sistemática la calidad de los *Prompts* y desarrollar estrategias que apoyen la generación de entradas más claras y estructuradas.

1.3 Complejidad del Problema

La investigación enfrenta complejidad debido a la ambigüedad del lenguaje natural y las dificultades que presentan los usuarios no expertos al diseñar *Prompts* efectivos para los LLMs. La falta de experiencia y la ausencia de métricas estandarizadas en la literatura para evaluar textos cortos o *Prompts* dificulta la evaluación, comparación y mejora de las entradas generadas por los usuarios no expertos.

De acuerdo con (Zamfirescu-Pereira et al., 2023), las personas no expertas diseñan *Prompts* siguiendo expectativas de interacción humana, lo que provoca errores e inconsistencias y limita la efectividad de los LLMs.

Por otro lado, (Burlin, 2023) señala que los LLMs son complejos y difíciles de entender para los usuarios. La información sobre cómo funcionan y por qué generan ciertos resultados no es suficiente para los diseñadores sin experiencia, lo que dificulta que puedan colaborar efectivamente con estos LLMs y aprovecharlos de manera creativa. Además, los usuarios deben aprender a adaptar su forma de trabajar y comunicarse con los LLMs para obtener resultados útiles, ya que la falta de explicabilidad y transparencia limita su control y confianza sobre los resultados generados.

Finalmente, (Denny et al., 2023) establece que enseñar a construir *Prompts* requiere ejercicios iterativos y retroalimentación sistemática, fomentando habilidades de evaluación y optimización en los estudiantes.

1.4 Objetivos de la Tesis

1.4.1 Objetivo General

El objetivo general de esta tesis es evaluar y proponer estrategias para mejorar la calidad de los *Prompts* en español culto que sirven como entrada de un LLM, empleando conceptos de Ingeniería de *Prompts* y NLP.

1.4.2 Objetivos Específicos

- Implementar criterios y métricas de calidad de texto que permitan evaluar de manera sistemática los *Prompts*.
- Diseñar y desarrollar una técnica para optimizar *Prompts*, combinando evaluación cuantitativa y estrategias de Ingeniería de *Prompts* y NLP.
- Analizar cómo la aplicación de la técnica de mejora influye en la calidad lingüística y estructural de los *Prompts* antes de ser procesados por un LLM.
- Evaluar la efectividad de la técnica de mejora de *Prompts* mediante experimentos controlados con criterios definidos para selección de *Prompts* y LLMs.
- Desarrollar una aplicación web que apoye la generación de *Prompts* mejorados.

1.5 Alcances y Limitaciones

1.5.1 Alcances

- Explorar técnicas de Ingeniería de *Prompts*, actualmente utilizadas para mejorar la generación de texto en los LLMs, como base para crear una técnica de mejora de *Prompts*.
- Utilizar un corpus en español culto, acotado a un dominio específico.
- Desarrollar e implementar una técnica de mejora de *Prompts* mediante estrategias de Ingeniería de *Prompts* y NLP.
- Evaluar la técnica de mejora de *Prompts* mediante experimentos complementarios que incluyen: obtención y selección de *Prompts*, análisis de LLMs, cálculo de métricas de calidad y desempeño, determinación de parámetros de longitud mínima y máxima tanto de los *Prompts* como de las respuestas generadas por los LLMs.
- Analizar los cambios en los *Prompts* antes y después de aplicar la técnica, utilizando una métrica cuantitativa de calidad.
- La técnica de mejora de *Prompts* está enfocada exclusivamente en el idioma español culto.

1.5.2 Limitaciones

- Los LLMs utilizados no serán modificados.
- La generalización de los *Prompts* no se considera, dado que el estudio se centra en un dominio específico.
- No se analiza ni aborda el posible sesgo presente en los LLMs empleados.
- Se considera el costo y la disponibilidad de un corpus adecuado al dominio seleccionado.
- La evaluación de la técnica de mejora se limita a las métricas definidas dentro del estudio; no se garantiza que los resultados sean aplicables a todos los LLMs o dominios.

1.6 Análisis del Problema

Los LLMs han demostrado capacidades sobresalientes para ejecutar tareas complejas mediante instrucciones en lenguaje natural. Sin embargo, el desempeño de cada LLM depende fuertemente de la calidad de las instrucciones o *Prompts* que recibe. A continuación, se presentan los hallazgos principales de estudios recientes.

De acuerdo con (Chen et al., 2023), la calidad del *Prompt* es un factor determinante para la efectividad de un LLM. Instrucciones claras, bien estructuradas y con la longitud adecuada aumentan significativamente la precisión y consistencia de las respuestas, mientras que *Prompts* ambiguos o mal redactados limitan el rendimiento del LLM.

Por otro lado, (Park et al., 2023) muestra que la optimización de la redacción y estructura de los *Prompts* permite guiar un LLM hacia resultados más precisos y coherentes. Esto evidencia la importancia de aplicar métodos sistemáticos de selección y refinamiento de *Prompts*, especialmente para usuarios sin experiencia en su elaboración.

Finalmente, estos estudios evidencian la necesidad de métodos automatizados y estrategias de Ingeniería de *Prompts* para maximizar la efectividad de los LLMs, reduciendo la dependencia de conocimientos expertos y mejorando la calidad de las respuestas generadas.

1.7 Justificación

Esta investigación se justifica por la necesidad de mejorar la calidad de los *Prompts* utilizados como entrada en los LLMs y que son escritos por usuarios no expertos. Dado que un *Prompt*

puede contener ambigüedades, información incompleta o falta de contexto, se vuelve esencial analizar y optimizar su redacción para favorecer interpretaciones más claras y respuestas más coherentes por parte de los LLMs.

De acuerdo con (Velásquez-Henao et al., 2023a), la falta de claridad, especificidad o estructura adecuada en un *Prompt* puede conducir a respuestas vagas, imprecisas o contextualmente inapropiadas. Su estudio muestra que la construcción de instrucciones efectivas requiere un proceso iterativo que incluye definir el objetivo, diseñar el mensaje, evaluar la respuesta e introducir mejoras. Esta perspectiva respalda la importancia de estudiar cómo la calidad del *Prompt* influye directamente en el rendimiento de los LLMs.

Por otro lado, (Wang et al., 2023) señalan que la Ingeniería de *Prompts* es fundamental para guiar el comportamiento de los LLMs en tareas especializadas, especialmente en dominios donde la precisión contextual es indispensable. Su análisis evidencia que la estructura, claridad y suficiencia informativa de las instrucciones influyen significativamente en la capacidad del LLM para generar resultados útiles y adaptados al contexto.

En conjunto, ambos trabajos sustentan la relevancia de mejorar la redacción, estructura y contenido de los *Prompts*. Esta investigación se apoya en dichas contribuciones para analizar las características que influyen en el desempeño de los LLMs y para fortalecer la construcción de instrucciones más claras y efectivas, sentando bases para el desarrollo de futuras métricas de evaluación de *Prompts*.

1.8 Estructura del Documento

La tesis se organiza en cinco capítulos. El Capítulo 1 introduce el contexto, los objetivos, la justificación y la delimitación del problema; el Capítulo 2 aborda el marco teórico, revisa antecedentes y explora el estado del arte relacionado con el problema; el Capítulo 3 describe el diseño e implementación de la solución; el Capítulo 4 presenta los resultados experimentales; y el Capítulo 5 concluye, destacando logros, líneas futuras y productos generados. Además, se incluyen secciones de referencias y anexos.

Capítulo 2

Fundamentos Conceptuales y Análisis de Soluciones Previas

2.1 Conceptos Clave

2.1.1 Usuario Experto

De acuerdo con (Popovic, 2000), el conocimiento se entiende como un conjunto que integra saberes generales y específicos, experiencia en la realización de tareas y habilidades cognitivas avanzadas.

2.1.2 Calidad Textual

De acuerdo con (Cuberos, 2019), la calidad textual se logra mediante el uso efectivo de dispositivos léxicos, sintácticos y discursivos apropiados al contexto comunicativo.

2.1.3 *Prompt*

De acuerdo con (Wong et al., 2023), el texto se expresa en Lenguaje Natural de forma libre, aunque también puede adoptar la forma de imágenes de referencia que contienen pistas sobre conceptos o preferencias objetivo.

- **Ejemplo:** Si alguien menciona "*un día soleado en la playa*" en una conversación, esa es una forma de texto en Lenguaje Natural. Pero si muestra una foto de una playa con el sol brillando en el cielo, eso sería una imagen de referencia que ayuda a entender de qué están hablando.

2.1.4 Técnico

De acuerdo con (Diccionario de la lengua española, 2024), se considera técnico a la persona que posee conocimientos especializados en una ciencia o arte.

2.1.5 Ingeniería de *Prompts*

De acuerdo con (DAIR.AI, 2025), la Ingeniería de *Prompts* es una disciplina relativamente nueva dedicada al desarrollo y la optimización de *Prompts*, con el fin de utilizar eficientemente los LLMs en una amplia variedad de aplicaciones y temas de investigación, tales como:

- Mejorar la capacidad de los LLMs en una amplia gama de tareas comunes y complejas.
- Responder preguntas y razonamiento aritmético.
- Diseñar técnicas de *Prompts* robustas y efectivas que interactúen con los LLMs y otras herramientas.

Algunas de las técnicas propuestas para aplicar *Prompting* (proceso de diseñar y formular *Prompts* o instrucciones para los LLMs) son las siguientes:

- *Prompt* sin entrenamiento previo (*Zero-shot*).
- *Prompt* con pocas muestras (*Few-shot*).
- *Prompt* Cadena de razonamiento (del inglés *Chain-of-Thought* o CoT).
- Autoconsistencia.
- *Prompt* de conocimiento generado.
- Ingeniería de *Prompts* automática (del inglés *Automatic Prompt Engineering* o APE).
- *Prompt* activo.
- *Prompt* de estímulo direccional.
- *T*écnica de Razonamiento y Acción (del inglés *Reason Act* o ReAct)
- *Prompt* CoT multimodal.
- *Prompt* de grafo.

2.1.6 Procesamiento de Lenguaje Natural

Según (Liddy, 2001), el NLP es un rango de técnicas computacionales motivadas teóricamente para analizar y representar textos de ocurrencia natural en uno o más niveles de análisis lingüístico con el propósito de lograr un procesamiento del lenguaje similar al humano para una variedad de tareas o aplicaciones.

2.1.7 Ajuste Fino

De acuerdo con (Bergmann, 2024), el ajuste fin es un proceso mediante el cual un LLM pre-entrenado se adapta a tareas específicas o casos de uso, aprovechando el conocimiento que el modelo ya ha aprendido como punto de partida para aprender nuevas tareas.

2.1.8 Transformadores Generativos Pre-entrenados

De acuerdo con (Yenduri et al., 2023), los Transformadores Generativos Pre-entrenados (del inglés *Generative Pre-trained Transformer* o GPT) son LLMs preentrenados con grandes cantidades de datos textuales y capaces de realizar una amplia gama de tareas relacionadas con el Lenguaje Natural. Además, el autor comenta que:

"Un GPT es un LLM basado en Aprendizaje Profundo que puede generar textos similares a los humanos basados en una entrada textual dada".

2.1.9 Modelos de Lenguaje

De acuerdo con (Yue et al., 2015), los Modelos de Lenguaje (del inglés *Language Models*) o LMs) reducen, inducen y expresan las reglas del Lenguage Natural utilizadas por los seres humanos a través de la Modelización Matemática.

Entre los LMs se pueden identificar dos tipos principales:

- **Modelo de Lenguaje Gramatical.** Es un enfoque lingüístico que se basa en la estructura y reglas gramaticales para comprender y generar texto de manera coherente y correcta desde el punto de vista gramatical.
- **Modelo de Lenguaje Estadístico.** Es un enfoque basado en el análisis de la probabilidad y frecuencia de ocurrencia de palabras y secuencias de palabras en un corpus de texto, utilizado para predecir y generar texto en función de patrones estadísticos.

2.1.10 Modelos Grandes de Lenguaje

De acuerdo a (Zhao et al., 2023) los LLMs, también llamados Modelos de Lenguaje Transformer, tienen las siguientes características:

- Contienen cientos de miles de millones (o más) de parámetros.
- Están entrenados con grandes volúmenes de datos de texto. Entre estos modelos se encuentran GPT-5, GPT-4, GPT-3 y Galactica.
- Muestran una gran capacidad para comprender el lenguaje natural y resolver tareas complejas.

2.2 Métricas para Evaluar la Calidad del Texto

2.2.1 PromptIQ

Es una métrica propuesta en este trabajo que evalúa la calidad de los *Prompts*, considerando errores lingüísticos y estructurales que proporcionan una evaluación cuantitativa integral. La efectividad de PromptIQ se analiza en la Sección 4.4.1, a través de la evaluación de métricas complementarias como son: coherencia, relevancia y diversidad.

2.2.2 Coherencia y Cohesión

De acuerdo con (Candelo et al., 2018), la coherencia se refiere a la conexión de ideas que permite formar el significado principal del texto, considerándose así una característica esencial. (Graesser et al., 2004) indica que la coherencia constituye una característica de la representación mental del contenido del texto por parte del lector.

Por otro lado la cohesión se refiere a los aspectos retóricos necesarios para desarrollar argumentos de apoyo en el texto, y es una propiedad objetiva del lenguaje y el texto explícito (Graesser et al., 2004).

2.2.3 Gramaticalidad

De acuerdo con (Bakar et al., 2017), la gramática es un campo que se centra en la formación de palabras y en el proceso de construcción de oraciones en cualquier idioma. La gramática constituye un conjunto de reglas sobre cómo se forma una palabra en particular y cómo esa palabra se combina con otras para producir una oración gramatical.

Los lingüistas la definen como un conjunto de sistemas que deben ser obedecidos por un usuario del lenguaje, convirtiéndose en un concepto fundamental para la producción de un lenguaje correcto y estéticamente adecuado.

La gramática es algo obligatorio en la estructura del orden del discurso o la escritura y una clasificación.

2.2.4 Relevancia

Evalúa la capacidad de un texto para proporcionar información útil y significativa que facilite la comprensión y el aprendizaje en el contexto de un LM. Esta métrica se mide mediante enfoques cualitativos y cuantitativos, incluyendo el análisis de contenido para valorar la relevancia del material presentado. Una puntuación alta en esta métrica indica que el contenido es adecuado y valioso.

Por otro lado, según (Ramos et al., 2003), la relevancia de una palabra w en un documento d puede determinarse mediante la fórmula Frecuencia de Término – Frecuencia Inversa de Documento (del inglés *Term Frequency – Inverse Document Frequency* o TF-IDF), la cual combina la frecuencia de dicha palabra en el documento con su rareza en el corpus D, como se muestra en la Ecuación (2.1):

$$wd = f_{w,d} \cdot \log\left(\frac{|D|}{f_{w,D}}\right) \tag{2.1}$$

Donde:

- $f_{w,d}$ representa la frecuencia de la palabra w en el documento d.
- $|D|$ es el tamaño del corpus, es decir, el número total de documentos.
- $f_{w,D}$ indica la cantidad de documentos en los que aparece la palabra w.

2.2.5 Diversidad

Esta métrica mide la variabilidad en los textos, incluyendo el vocabulario, la estructura de las frases y las perspectivas.

De acuerdo con (Borden et al., 2022), el índice de Simpson es una medida de la diversidad de un atributo dentro de una población. Este índice se calcula determinando la probabilidad de que dos individuos seleccionados al azar de la población provengan de diferentes categorías dentro de un atributo específico. Mide cuán diversa es la distribución de categorías dentro de un conjunto dado.

El índice de Simpson se basa en la proporción de individuos en cada categoría del atributo. Valores más cercanos a 1 indican una alta diversidad, mientras que valores cercanos a 0 indican baja diversidad. Se calcula utilizando la Ecuación (2.2):

$$D = 1 - \sum_{i=1}^{n} p_i^2 \tag{2.2}$$

donde:

- p_i es la proporción de individuos en la i-ésima categoría del atributo.
- n es el número total de categorías dentro del atributo de diversidad que se está midiendo.

Se toma como referencia la definición propuesta por (Borden et al., 2022), debido a la relación entre el concepto de “individuo” en la población y el de “texto” dentro de un corpus de documentos. Aunque el índice de Simpson se utiliza principalmente para medir la diversidad en poblaciones, su aplicación ha sido extendida al análisis de datos textuales, como en la evaluación de la diversidad temática dentro de un corpus de documentos.

2.3 Errores Lingüísticos

2.3.1 Contexto

Son los errores que afectan la interpretación debido a factores situacionales, como el contexto cultural o la intención comunicativa. La falta de consideración del contexto puede generar ambigüedades y malentendidos.

2.3.2 Fluidez

Se refiere a la ausencia de errores que interrumpen la legibilidad y el ritmo del texto. La falta de fluidez genera confusión y dificulta el seguimiento del contenido, afectando la comprensión general.

2.3.3 Gramática

Conjunto de reglas que regulan el uso correcto de los tiempos verbales, la concordancia y las preposiciones. Los errores gramaticales afectan la precisión y formalidad del lenguaje, disminuyendo la calidad del texto.

2.3.4 Incoherencia

Errores en la conexión lógica entre ideas. Las incoherencias pueden causar confusión en el lector y dificultar la comprensión del mensaje.

2.3.5 Léxico

Se refiere a los errores en el uso del vocabulario, ya sea por limitaciones o por empleo inadecuado de términos. Un léxico inapropiado o limitado disminuye la riqueza y efectividad del contenido.

2.3.6 Ortografía

Conjunto de normas y convenciones que regulan la correcta escritura de las palabras en un idioma. Los errores ortográficos consisten en fallos al escribir que pueden dificultar la comprensión del mensaje, afectando la precisión y claridad del texto, y comprometiendo su correcta interpretación.

2.3.7 Pragmatismo

Se define como la capacidad del texto para emplear el lenguaje de manera adecuada según el propósito comunicativo y el contexto en el que se produce. Esta variable se evalúa mediante

un conteo discreto de errores relacionados con el uso práctico del lenguaje en situaciones específicas. Una baja puntuación en esta métrica indica dificultades para adaptar el discurso a las necesidades comunicativas del entorno.

2.3.8 Puntuación

Se define como la cantidad de errores en el uso de signos de puntuación que estructuran el texto, tales como comas, puntos y signos de interrogación. Estos errores afectan la claridad y fluidez del texto, modificando la interpretación y comprensión de las ideas expresadas. Cada error individual contribuye al total acumulado, lo que permite una evaluación cuantitativa de la precisión en el uso de la puntuación dentro del análisis del texto.

2.3.9 Registro

Se refiere a los errores en la adecuación del lenguaje según el nivel de formalidad requerido. El uso inadecuado del registro afecta la efectividad del mensaje en función del contexto y la audiencia.

2.3.10 Semántica

Son los errores en el significado de palabras o frases. Los errores semánticos pueden provocar ambigüedades o malentendidos, afectando la claridad y precisión del mensaje.

2.3.11 Sintaxis

Se refiere a la estructura de las oraciones y al orden correcto de las palabras. Los problemas sintácticos afectan la coherencia y claridad del texto, dificultando la comprensión de las ideas.

2.4 Métricas de Evaluación de Modelos de Lenguaje

2.4.1 Perplejidad

Según (Zhang et al., 2018), en el modelado de lenguaje probabilístico, la perplejidad es un método de evaluación clásico para el modelo de temas. La idea básica de la perplejidad es que el LM otorga un valor de probabilidad más alto al conjunto de prueba.

Un grado menor de perplejidad significa que el modelo tendría un impacto efectivo en nuevos textos. Por lo tanto, el grado de confusión generalmente disminuye con el aumento del número de posibles temas. Para el modelo de temas, la Ecuación (2.3) se aplica para calcular la perplejidad de la siguiente manera:

$$\text{Perplejidad} = b^{-\frac{1}{N}\sum_{k=1}^{N} \log_b p(w_k)} \tag{2.3}$$

donde:

- b es la base del logaritmo (generalmente $b = e$ o $b = 2$),
- N es el número total de palabras en el conjunto de datos,
- $p(w_k)$ es la probabilidad del k-ésimo término, calculada como el producto de la distribución de palabras y la distribución de temas del texto.

Por otro lado, la perplejidad de una secuencia X puede expresarse como:

$$\text{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i=1}^{t} \log p_\theta(x_i \mid x_{<i})\right\} \tag{2.4}$$

donde:

- t es el número de tokens en la secuencia X,
- x_i es el i-ésimo token de la secuencia,
- $x_{<i}$ representa todos los tokens previos a x_i,
- θ son los parámetros del modelo pre-entrenado,
- $p_\theta(x_i \mid x_{<i})$ es la probabilidad de que el token x_i ocurra dado el historial de tokens $x_{<i}$ y los parámetros θ.

Esta definición permite interpretar la perplejidad como una medida de incertidumbre promedio que tiene un modelo al predecir cada token de una secuencia.

El Algoritmo 1 implementa este cálculo de manera práctica, considerando la longitud de ventana de procesamiento, la segmentación de la secuencia y la exclusión de posiciones irrelevantes mediante el valor -100, mientras que $outputs.loss$ representa la pérdida logarítmica negativa promedio para cada rango de tokens.

Algoritmo 1: Calcular Perplejidad

Input: STRIDE: longitud de la ventana de procedimiento, M_LEN: longitud máxima de la ventana de procesamiento, dependiente del Modelo Pre-entrenado GPT-2, model: el modelo pre-entrenado GPT-2, tokenizer: el tokenizador del Modelo pre-entrenado GPT-2, text: el texto a analizar
Output: Perplejidad del texto analizado
Function `CALCULAR PERPLEJIDAD(`*text*`)`**:**

```
seq_len, encodings ← tokenizer(text) ;  // obtener una serie de representaciones de
  texto
nlls ← [ ] ;              // probabilidades logarítmicas negativas - lista vacía
prev_end_loc ← 0;
end_loc ← 0;
begin_loc ← 0;
while begin_loc ≤ seq_len do
    end_loc ← min(begin_loc + M_LEN, seq_len);
    trg_len ← end_loc - prev_end_loc;
    input_ids ← encodings[begin_loc:end_loc] ;     // recuperar la representación de
      texto en el espacio del Modelo para un rango
    target_ids ← clone(input_ids);
    target_ids[0:array length - trg_len] ← -100 ; // excluir los valores específicos de
      los cálculos siguientes, marcándolos así con -100
    outputs ← model(target_ids) ;         // resultado para cada token en el rango
      seleccionado
    nlls.append(outputs.loss);
    if end_loc == seq_len then
        break;
    end
    prev_end_loc ← end_loc;
    begin_loc ← begin_loc + M_LEN;
end
return e^{mean(nlls)};
end;
```

2.4.2 *Bilingual Evaluation Understudy*

De acuerdo a (Rosario & Noever, 2023), el puntaje Evaluador Bilingüe Automático (del inglés *Bilingual Evaluation Understudy* o BLEU) se desarrolló inicialmente para medir el rendimiento de la capacidad de una máquina para traducir texto de un idioma a otro manteniendo un contexto y significado apropiados, de ahí el título bilingüe. BLEU puede comparar la traducción de una máquina, también llamada traducción candidata, con una traducción humana existente, conocida como traducción de referencia.

El algoritmo para calcular el puntaje BLEU funciona mediante la tokenización de la candidata y el cálculo de la precisión en función de cuántas palabras aparecen en la traducción candidata que también aparecieron en la traducción de referencia, también llamada precisión.

De acuerdo con (Noorbehbahani & Kardan, 2011), la métrica de BLEU se calcula mediante la

Ecuación (2.5):

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{2.5}$$

donde:

- BP: Penalización por brevedad (del inglés *Brevity Penalty* o BP), un factor que reduce la puntuación cuando la traducción es más corta que la referencia.
- N: número máximo de n-gramas considerados en la evaluación (por ejemplo, $N = 4$ para BLEU-4).
- w_n: peso asignado a cada n-grama, con $w_n \in [0, 1]$ y $\sum_{n=1}^{N} w_n = 1$.
- p_n: precisión modificada de los n-gramas, definida como la proporción de n-gramas en la traducción que coinciden con los n-gramas de referencia (ver Ecuación (2.6)).
- $\exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$: media geométrica ponderada de las precisiones de los n-gramas.

La precisión modificada p_n se calcula como:

$$p_n = \frac{\sum_{C \in \{candidates\}} \sum_{n\text{-gram} \in C} \text{Count}_{clip}(n\text{-gram})}{\sum_{C' \in \{candidates\}} \sum_{n\text{-gram}' \in C'} \text{Count}(n\text{-gram}')} \tag{2.6}$$

donde:

- $\{candidates\}$: conjunto de traducciones generadas automáticamente.
- $\text{Count}_{clip}(n\text{-gram})$: número de coincidencias de un n-grama en la traducción que no exceda su frecuencia en la referencia.
- $\text{Count}(n\text{-gram})$: número total de ocurrencias de cada n-grama en la traducción.

El factor de penalización por brevedad se define como:

$$\text{BP} = \begin{cases} 1 & \text{si } C > r \\ e^{\left(1 - \frac{r}{C}\right)} & \text{si } C \leq r \end{cases} \tag{2.7}$$

donde:

- C: longitud total de la traducción candidata.
- r: longitud total de la traducción de referencia.

Alternativamente, se puede expresar BLEU en términos de logaritmos como:

$$\log \text{BLEU} = \text{mín}\left(1 - \frac{r}{C}, 0\right) + \sum_{n=1}^{N} w_n \log p_n \tag{2.8}$$

Para el enfoque básico de BLEU, se utiliza $N = 4$ y pesos uniformes $w_n = \frac{1}{N}$ para todos los n-gramas.

2.4.3 *Recall-Oriented Understudy for Gisting Evaluation*

De acuerdo a (Rasheed Issam, Patel et al., 2021), la Evaluación de Resumen Orientada al Recuerdo (del inglés *Recall-Oriented Understudy for Gisting Evaluation* o ROUGE) es el conjunto de métricas más popular utilizado para evaluar resúmenes generados por máquinas. Compara los resúmenes base (resúmenes de referencia o resúmenes verdaderos) presentes en el conjunto de datos fuente con los resúmenes generados automáticamente por modelos de máquinas. Los resúmenes base suelen ser escritos por seres humanos.

El conjunto de métricas ROUGE presenta cinco métricas de evaluación diferentes. ROUGE-N (ROUGE-1, ROUGE-2 y así sucesivamente), que miden la superposición entre n-gramas del resumen base y los del resumen generado automáticamente.

Por otro lado, (Ganesan, 2018) comenta que:

> **"Las puntuaciones de ROUGE no reflejan la verdadera calidad de los resúmenes y limitan la evaluación multifacética de los mismos"**

Por lo cual propone ROUGE 2.0, altamente modular, permitiendo la sustitución del diccionario de sinónimos por diccionarios específicos de dominio o idioma, lo que facilita su adaptación a tareas de resumen en diferentes contextos, como *Twitter* o lenguajes específicos.

ROUGE-*Topic* permite especificar qué combinaciones de categorías gramaticales POS se deben utilizar para la evaluación mediante uni-gramas. Mediante las Ecuaciones (2.9) y (2.10), se pueden obtener los valores de ROUGE-*Topic* desde dos enfoques distintos: *recall* y *precision*.

$$\text{ROUGE-}\textit{Topic}_{\textit{recall}} = \frac{\sum \text{Overlap}(\text{REF}_{i_{pos}}, \text{SYS}_{j_{pos}})}{|\text{REF}_{i_{pos}}|} \tag{2.9}$$

$$\text{ROUGE-}\textit{Topic}_{\textit{precision}} = \frac{\sum \text{Overlap}(\text{REF}_{i_{pos}}, \text{SYS}_{j_{pos}})}{|\text{SYS}_{j_{pos}}|} \tag{2.10}$$

Donde:

- $\text{REF}_{i_{pos}}$: conjunto de tokens de la categoría gramatical pos en el i-ésimo resumen de referencia.

- SYS$_{j_{pos}}$: conjunto de tokens de la categoría gramatical pos en el j-ésimo resumen generado por el sistema.
- pos: categoría gramatical (*noun*, *verb*, etc.) seleccionada para la evaluación.
- Overlap$(\text{REF}_{i_{pos}}, \text{SYS}_{j_{pos}})$: número de tokens comunes entre el resumen de referencia y el del sistema, considerando solo la categoría pos.

En muchos casos, las palabras relacionadas con el tema pueden repetirse tanto en los resúmenes de referencia como en los resúmenes generados por el sistema, por lo que este enfoque permite evaluar la cobertura de términos relevantes por categoría gramatical.

2.4.4 $F_1 - Score$

Deacuerdo a (Sitarz, 2022) esta métrica se utiliza comúnmente para evaluar el rendimiento de clasificadores binarios de aprendizaje automático. Calculada como una media armónica de precisión y recuperación, obtiene una ventaja sobre métricas menos complejas como la precisión, especialmente cuando se utiliza en conjuntos de datos desequilibrados.

De acuerdo a (Tatbul et al., 2018), *Precision* y *Recall* son esenciales para determinar el valor representativo de $F_1 - Score$, por lo cual sus valores se calculan mediante las Ecuaciones (2.11) y (2.12).

$$\textit{Precision} = \frac{TP}{TP + FP} \tag{2.11}$$

$$\textit{Recall} = \frac{TP}{TP + FN} \tag{2.12}$$

Donde:

- **Verdaderos positivos (del inglés *True Positives* o TP)**: número de instancias correctamente identificadas como positivas.
- **Falsos positivos (del inglés *False Positives* o FP)**: número de instancias incorrectamente identificadas como positivas.
- **Falsos negativos (del inglés *False Negatives* o FN)**: número de instancias que son positivas pero no fueron detectadas.

De manera informal, *Precision* representa la proporción de detecciones positivas que son realmente correctas, mientras que *Recall* indica la proporción de instancias positivas reales que fueron correctamente detectadas.

2.5 Métodos Estadísticos y de Visualización

2.5.1 Regresión Lineal Multivariable

La regresión lineal multivariable es una técnica estadística utilizada para modelar la relación entre una variable dependiente (o variable de respuesta) y múltiples variables independientes (o predictoras).

Ecuación del modelo:

El modelo de regresión lineal multivariable se expresa como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{2.13}$$

Donde:

- Y: variable dependiente o respuesta.
- β_0: término constante o intersección; valor de Y cuando todas las X_i son cero.
- $\beta_1, \ldots, \beta_n$: coeficientes de regresión; indican cuánto cambia Y al variar una unidad en X_i, manteniendo las demás variables constantes.
- $X_1, \ldots, X_n$: variables independientes o predictoras.
- ε: término de error, que captura la variación no explicada por las variables independientes.

2.5.2 Shapiro-Wilk *Test*

El *test* de Shapiro-Wilk es una prueba estadística utilizada para evaluar si un conjunto de datos sigue una distribución normal.

Hipótesis:

- **Hipótesis nula (H_0):** los datos provienen de una población con distribución normal.
- **Hipótesis alternativa (H_1):** los datos no provienen de una población con distribución normal.

El estadístico de Shapiro-Wilk se calcula comparando los datos observados con los valores esperados bajo normalidad. Si el p-valor es menor que un nivel de significancia $\alpha = 0.05$, se rechaza H_0, indicando que los datos no siguen una distribución normal.

2.5.3 Wilcoxon *Signed-Rank Test*

Es una prueba estadística no paramétrica utilizada para comparar dos muestras relacionadas, como mediciones antes y después de un tratamiento. A diferencia del *test* t de muestras pareadas, no requiere que los datos sigan una distribución normal, por lo que se usa cuando la normalidad no puede asumirse (por ejemplo, cuando en el *test* de Shapiro-Wilk los datos no son normales).

Hipótesis:

- **Hipótesis nula (H_0):** Las diferencias entre las muestras tienen mediana igual a cero (no hay efecto significativo).
- **Hipótesis alternativa (H_1):** Las diferencias entre las muestras tienen mediana distinta de cero (existe un efecto significativo).

El *test* de Wilcoxon se basa en los rangos de las diferencias absolutas y sus signos. Se rechaza H_0 si el valor p es menor que el nivel de significancia $\alpha = 0.05$.

2.5.4 Análisis de Componentes Principales

El Análisis de Componentes Principales (del inglés *Principal Component Analysis* o PCA) es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos multivariado, preservando la mayor cantidad posible de su variabilidad (información).

Consiste en transformar el conjunto original de variables, que pueden estar correlacionadas, en un nuevo conjunto de variables no correlacionadas llamadas componentes principales. Estas nuevas variables son combinaciones lineales de las originales y se ordenan de manera que:

- El primer componente principal explica la mayor parte de la varianza del conjunto de datos.
- El segundo componente principal explica la mayor cantidad de varianza posible del remanente, y así sucesivamente.

2.5.5 Centroide $k - means$

Según (Ahmed et al., 2020), el algoritmo de $k - means$ es un método de aprendizaje no supervisado utilizado para organizar datos no etiquetados en clústeres coherentes.

Funciona asignando cada punto de datos al clúster cuyo centroide está más cercano, generalmente usando la distancia euclidiana como medida de proximidad.

El objetivo del algoritmo es minimizar la suma de las distancias entre los puntos de datos y los centroides de sus clústeres asignados. Un mayor valor de k genera más clústeres pequeños, capturando mayor detalle, mientras que un menor valor de k produce clústeres más grandes con menor detalle.

2.5.6 Estimación de Densidad de *Kernel*

La Estimación de Densidad de *Kernel* (del inglés *Kernel Density Estimation* o KDE) es un método no paramétrico para estimar la función de densidad de probabilidad de una variable aleatoria a partir de un conjunto de datos.

En términos simples, KDE construye una curva suave que representa la distribución de los datos, colocando una función llamada *kernel* (por lo general una función gaussiana) centrada en cada punto de datos, y luego sumando todas esas funciones para obtener una estimación continua de la densidad de probabilidad.

Esto permite obtener una aproximación suave y flexible de la distribución sin asumir una forma específica (como una distribución normal), a diferencia de los histogramas que dependen de la elección de los intervalos o $bins$.

2.5.7 Histograma de Frecuencias

Es una representación gráfica que muestra cómo se distribuyen los datos en diferentes intervalos o clases (llamados $bins$). Cada barra del histograma representa la cantidad de datos (frecuencia) que cae dentro de un intervalo específico.

El eje horizontal indica los valores o rangos de la variable, mientras que el eje vertical muestra la frecuencia de datos en cada intervalo. El histograma es útil para visualizar la forma, dispersión y concentración de un conjunto de datos.

2.5.8 Gráfico de Líneas Multivariado

Un gráfico de líneas multivariado es una representación gráfica que permite visualizar simultáneamente múltiples variables cuantitativas, cada una mediante una línea distinta dentro de un mismo sistema de coordenadas.

El eje horizontal generalmente representa una variable categórica u ordinal (por ejemplo, periodos de tiempo o grupos de estudio), mientras que el eje vertical indica los valores cuantitativos correspondientes a cada variable.

Este tipo de gráfico resulta especialmente útil para comparar el comportamiento relativo de diferentes métricas o indicadores a lo largo de diversas categorías, facilitando la identificación de patrones, tendencias, interacciones y posibles divergencias entre las variables analizadas.

2.6 Exploración del Estado del Arte

En esta sección se presentan estudios que exploran técnicas aplicadas al pre-procesamiento de *Prompts*, en el contexto del estado del arte de esta tesis. Sin embargo, persiste la incertidumbre respecto a si existen actualmente trabajos que enfoquen su investigación específicamente en el pre-procesamiento de *Prompts*.

Esta situación abre la puerta a la exploración de otras vías de investigación que podrían contribuir en esta área, como el caso de los LLMs que emplean técnicas de procesamiento de texto.

En este contexto, se plantea la posibilidad de considerar dichas técnicas de procesamiento de texto aplicadas a los LLMs como una estrategia alternativa para abordar los desafíos del pre-procesamiento de *Prompts*.

A continuación, se describen trabajos de investigación publicados en los últimos cinco años (Agosto 2020 – Agosto 2025) relacionados con el tema de esta tesis. La lista incluye investigaciones centradas en un concepto emergente denominado Ingeniería de *Prompts*.

Este enfoque propone la aplicación de técnicas de procesamiento de texto en los LLMs, permitiéndoles enfrentar deficiencias en sus respuestas de salida. Dichas respuestas son evaluadas con métricas asociadas a la calidad del texto generado, actualmente utilizadas en el NLP.

Asimismo, se abordan temas sobre el pre-procesamiento de *Prompts* y el NLP, los cuales contribuirán a generar una base sólida de conocimiento para desarrollar una técnica que permita un pre-procesamiento mejorado de *Prompts* utilizados como entrada para los LLMs.

Los estudios explorados se organizarán en cuatro núcleos temáticos: NLP en aplicaciones específicas, similitud semántica en textos cortos, modelado de temas y LLMs pre-entrenados y sus aplicaciones.

2.6.1 Modelado de Temas

2.6.1.1 *A Comprehensive Empirical Study of Bias Mitigation Methods for LMs Classifiers* (Chen et al., 2023)

El artículo estudia el sesgo en software y su mitigación en LMs, evaluando 17 métodos con 11 métricas de rendimiento, 4 de equidad y 20 de compensación en 8 tareas. Los métodos

reducen el rendimiento en 42 % – 66 %, mejoran equidad en 24 % – 59 % y muestran compromisos en 25 % de escenarios. No hay método óptimo universal, pero uno es superior en 30 % de casos.

2.6.1.2 *OCTIS: Comparing and optimizing topic models is simple!* (Terragni et al., 2021)

Optimización de Comparación de Modelos de Temas de Forma Simple (del inglés *Optimizing Comparisons of Topic models Is Simple* o OCTIS) es un marco para entrenar y comparar modelos de temas que usa optimización bayesiana para ajustar hiperparámetros. Integra modelos clásicos y neuronales, y evalúa coherencia, diversidad, significancia y clasificación. Comparado con bibliotecas como Gensim y Herramienta para Aprendizaje Automático de Lenguaje (del inglés *Machine Learning for LanguagE Toolkit* o MALLET), OCTIS destaca por su interfaz web y ajuste bayesiano, sin definir una opción superior definitiva.

2.6.1.3 *The evolution of topic modeling* (Churchill & Singh, 2022)

El artículo analiza la evolución de los modelos de temas como Asignación Latente de Dirichlet (del inglés *Latent Dirichlet Allocation* o LDA), Factorización de Matrices No Negativas (del inglés *Non-negative Matrix Factorization* o NMF) y Modelo de Temas con Autoencoder Variacional (del inglés *Variational Autoencoder Topic Model* o VAETM), usando aprendizaje automático y redes neuronales para mejorar la coherencia y diversidad temática. Evalúa coherencia, diversidad y desempeño en identificación de tópicos relevantes. Destaca factores para elegir modelos según tamaño y tipo de datos, nivel de ruido, fuente, y tareas posteriores como clasificación y sistemas conversacionales.

2.6.1.4 *Topic modeling algorithms and applications: A survey* (Abdelrazek et al., 2023)

El artículo analiza la evolución de modelos de temas como LDA, NMF y VAETM, aplicando aprendizaje automático y redes neuronales para mejorar coherencia y diversidad temática. Se destacan factores clave para elegir modelos, considerando tamaño y ruido del conjunto de datos, tipo de documentos y tareas posteriores. Se concluye que adaptar modelos a la complejidad y contexto mejora la identificación de tópicos relevantes.

2.6.2 Modelos de Lenguaje Pre-entrenados y sus Aplicaciones

2.6.2.1 *How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven* (McCoy et al., 2023)

El estudio presenta la métrica Acuerdo Ordenado de la Variabilidad en Novedades de Ejemplos (del inglés *Ranked Agreement of Variability in Example Novelties* o RAVEN), diseñada

para evaluar la novedad en textos generados por LLMs. Los autores aplican RAVEN a GPT-2 en cuatro tamaños diferentes y encuentran que, aunque el modelo produce estructuras nuevas, también reproduce fragmentos del conjunto de entrenamiento. Asimismo, muestran que GPT-2 emplea generalización composicional y analógica, pero con ciertas limitaciones semánticas. Finalmente, concluyen que el tamaño del modelo no determina la novedad de los textos, según el análisis de duplicación basado en el tamaño de los n-gramas.

2.6.2.2 *Impact of word embedding models on text analytics in deep learning environment: a review* (Asudani et al., 2023)

El artículo compara modelos de aprendizaje profundo y técnicas de incrustación de palabras para análisis de texto. Concluye que las incrustaciones específicas del dominio y Memoria a Largo y Corto Plazo (del inglés *Long Short-Term Memory* o LSTM) mejoran el rendimiento. LSTM con *Word2Vec* alcanzó 99.55 % en precisión, seguido de Red Neuronal Convolucional (del inglés *Convolutional Neural Network* o CNN) con *Word2Vec* chino con 96.60 %. Representaciones Codificadas Bidireccionales a partir de Transformadores (del inglés *Bidirectional Encoder Representations from Transformers* o BERT) logró 95.00 % en precisión, y *fastText* obtuvo 91.00 % en $F1$-*score*, destacando su efectividad.

2.6.2.3 *Instance-aware Prompt Learning for Language Understanding and Generation* (Jin et al., 2023)

El artículo propone Aprendizaje de *Prompts* Sensibles a la Instancia (del inglés *Instance-Prompt Learning* o IPL), una técnica que asigna *Prompts* únicos a cada instancia para comprensión y generación de lenguaje. Comparado con el Ajsute Fino (del inglés *Fine-tuning* o FT) y Ajuste de *Prompts* (del inglés *Prompt-tuning* o PT), IPL se aplica a GPT2-*base* y GPT2-*large* y se evalúa con múltiples métricas. IPL obtiene mejores resultados en Puntaje BERT*Score* 95.43 % y 95.47 %, BLEU 69.82 % y 68.53 %, Instituto Nacional de Estándares y Tecnología (del inglés *National Institute of Standards and Technology* o NIST) 8.82 % y 8.68 %, ROUGE-L 71.65 % y 71.20 %, Evaluación de Descripción de Imagen Basada en Consenso (del inglés *Consensus-based Image Description Evaluation* o CIDEr) 2.49 % y 2.51 %, Métrica para Evaluación de Traducción con Orden Explícito (del inglés *Metric for Evaluation of Translation with Explicit ORdering* o METEOR) 72.23 % y 50.55 % y Tasa de Edición de Traducción (del inglés *Translation Edit Rate* o TER) 48.50 % y 46.17 %, superando a FT y PT en rendimiento general. Las métricas empleadas se expresan en escalas numéricas donde valores más altos indican mejor desempeño, excepto TER, donde valores menores son preferibles.

2.6.2.4 *Interactive and visual Prompt engineering for ad-hoc task adaptation with large language models* (Strobelt et al., 2022)

El artículo presenta *PromptIDE*, un sistema interactivo para personalizar modelos de NLP sin experiencia en entrenamiento. Facilita la exploración y refinamiento de *Prompts* con visualizaciones y métricas. En la evaluación, *PromptIDE* obtuvo 84 respuestas correctas y 16 incorrec-

tas. La clasificación mostró 29 aciertos en política mundial, 28 en deportes, 13 en negocios y 14 en ciencia y tecnología.

2.6.2.5 *Knowledge enhanced Prompt tuning for Stance Detection* (Huang et al., 2023)

El artículo aborda la detección automática de postura en textos usando técnicas de NLP y aprendizaje automático. Se aplican métodos tradicionales y aprendizaje profundo en *datasets* de *tweets* en inglés y *microblogs* en chino y español. Se evalúa el rendimiento con métricas $F1$-*score*, *precision*, *recall* y exactitud, expresadas en valores numéricos. La creación de *datasets* anotados multilingües es fundamental.

2.6.2.6 *Measuring and Improving Consistency in Pretrained Language Models* (Elazar et al., 2021)

El estudio evalúa la consistencia en la extracción de conocimiento utilizando *PARAREL*, una técnica basada en 328 patrones que describen 38 relaciones semánticas. Se prueban Modelos de Lenguaje Pre-entrenados (del inglés *Pretrained Language Models* o PLMs) como BERT, RoBERTa y ALBERT. Se emplean métricas *Succ-Patt*, *Succ-Objs*, *Unk-Const*, *Know-Const*, *Accuracy*, *Consistency* y *Consistent-Acc*, reportadas en porcentajes con su respectiva desviación estándar. BERT-*large-wwm* destacó con *Succ-Patt* de 100.0 % ± 0.0, *Accuracy* de 48.7 % ± 25.0, y *Consistency* de 60.9 % ± 24.2.

2.6.2.7 *NIR-Prompt: A Multi-task Generalized Neural Information Retrieval Training Framework* (Xu et al., 2023)

El artículo presenta Recuperación de Información Neuronal - *Prompt* (del inglés *Neural Information Retrieval - Prompt* o NIR-*Prompt*), una técnica basada en BERT-*base* (109M) que fija *tokens* de indicación durante el entrenamiento para mejorar la recuperación de información. Se evaluó con métricas *Precision*, *Accuracy* y $F1$-*score* en tareas de Paráfrasis (del inglés *Paraphrase Identification* o PI) e Inferencia de Lenguaje Natural (del inglés *Natural Language Inference* o NLI). El mejor rendimiento se obtuvo con $k = 11$, logrando un *Accuracy* de hasta 79.8 % y $F1$-*score* de 83.8 % en PI, y 77.5 % y 84.5 % en NLI, respectivamente.

2.6.2.8 *Personalized Prompt Learning for Explainable* Recommendation (L. Li et al., 2023)

El estudio mejora la explicabilidad en sistemas de recomendación mediante Aprendizaje de *Prompt's* Personalizados para Recomendación Explicable – *Prompt's* Discretos (del inglés *Personalized Prompt Learning for Explainable Recommendation – Discrete Prompts* o PEPLER-D, un método que utiliza PLMs como GPT-2 y aplica estrategias como *Sequential Tuning*,

Prompt+LM, FT y *Fixed-M* PT. Se evalúa con *BLEU-4* y se compara con *BERT* y GPT-3. PEPLER-D supera a estos modelos, alcanzando *BLEU-4* de hasta 0.96 en *Amazon*, 0.9 en *TripAdvisor* y 0.69 en *Yelp*, con *learning rates* de 10^{-5} a 10^{-3}.

2.6.2.9 *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing* (Liu et al., 2023)

El artículo analiza el aprendizaje basado en *Prompts* aplicando NLP, que modela la probabilidad del texto mediante *Prompts* para llenar información incompleta. Permite pre-entrenar modelos con gran texto sin procesar y adaptarlos con pocos ejemplos. Se organizan modelos, *Prompts* y ajustes. Entre 2018-2021, aumentaron investigaciones, con 29 en 2021. Las tareas incluyen clasificación de texto (22), sondeo factual (15) y generación condicional (8).

2.6.2.10 *Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering* (Velásquez-Henao et al., 2023b)

El artículo propone la metodología Ingeniería Iterativa de *Prompts* Generativos (del inglés *Generative Prompt Engineering Iterative* o GPEI) para optimizar la interacción con LLMs con Inteligencia Artificial en ingeniería, mediante el diseño iterativo de *Prompts* precisos y contextuales. GPEI integra principios de Inteligencia Artificial explicable y evaluación continua de respuestas, regresando al diseño del *Prompt* si la respuesta no es satisfactoria, promoviendo respuestas transparentes y justificables para mejorar la comunicación con modelos Inteligencia Artificial.

2.6.2.11 *Recent advances in natural language processing via large pre-trained language models: A survey* (Min et al., 2023)

El artículo presenta la técnica de tareas *proxy* estilo preguntas y respuestas para entrenar modelos como BERT y RoBERTa-*large*. Esta técnica usa plantillas con contexto "C", *Prompt* "Q" y respuesta "A". Se aplica en clasificación de emociones y extracción de argumentos en eventos, destacando esta última para generar respuestas precisas basadas en el contexto proporcionado.

2.6.3 Aplicaciones de NLP

2.6.3.1 *From Natural Language Processing to Neural Databases* (Thorne et al., 2021)

Este estudio presenta *NeuralDB*, un sistema basado en modelos de transformadores de NLP que permite realizar consultas en lenguaje natural sin necesidad de un esquema de base de datos predefinido. *NeuralDB* logra una precisión del 79.45 % en coincidencia exacta para consultas que cuentan elementos y un 100 % en consultas que buscan valores mínimos o

máximos (*Min/Max Sets*). Estos resultados son mejores que los obtenidos por métodos que combinan Frecuencia Inversa de Término en Documentos (del inglés *Term Frequency – Inverse Document Frequency* o TF-IDF) con el modelo Transformador de Texto a Texto (del inglés *Text-to-Text Transfer Transformer* o T5) y el recuperador denso de pasajes junto con T5, los cuales alcanzaron 31.06 % y 38.07 % respectivamente. La métrica utilizada es el porcentaje de coincidencia exacta, que mide qué tan precisas son las respuestas del sistema.

2.6.3.2 *Improving patient self-description in Chinese online consultation using contextual Prompts* (X. Li et al., 2022)

El artículo presenta un modelo de aprendizaje automático para consultas médicas en línea, evaluado con métricas de *accuracy*, $F1$-*score macro*, Área bajo la curva (del inglés *Area Under the Curve* o AUC) *macro*, Coeficiente de correlación de Matthews (del inglés *Matthews Correlation Coefficient* o MCC) *macro* y entropía en tres corpus clínicos: pediatría, andrología y cardiología. El modelo con *Prompts* aprendidos mejora el rendimiento sobre heurísticas, con *precision* entre 0.33 y 0.70, AUC de 0.80 a 0.92 y MCC hasta 0.56. Estrategias basadas en certeza e incertidumbre no aportan mejoras significativas.

2.6.3.3 *Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking* (Y.-C. Zhou et al., 2022)

El artículo presenta un método para la interpretación automatizada de reglas en documentos regulatorios mediante etiquetado semántico y análisis sintáctico. Utiliza aprendizaje profundo para generar reglas ejecutables y mejorar la comprensión de requisitos complejos. Los resultados muestran una *precision* promedio del 86.2 %, *recall* del 86.3 % y $F1$-*score* de 86.2 % en un conjunto de validación con 4,575 etiquetas.

2.6.3.4 *Natural language processing for Nepali text: a review* (Shahi & Sitaula, 2022)

El artículo analiza el (NLP) para Nepalí, destacando métodos de aprendizaje automático y la creación de recursos lingüísticos específicos. Se proponen estrategias como la transferencia de conocimiento usando modelos preentrenados y el desarrollo de grandes conjuntos de datos anotados para mejorar la precisión. Las tareas incluyen clasificación con el Clasificador Bayesiano Ingenuo (del inglés *Naive Bayes* o NB), Máquina de Vectores de Soporte (del inglés *Support Vector Machine* o SVM), Red Neuronal Artificial (del inglés *Artificial Neural Network* o ANN) y Memoria a Largo y Corto Plazo (del inglés *Long Short-Term Memory* o LSTM); agrupación con *K-means* y *Fuzzy Set*; análisis de sentimientos con *Logistic Regression* (LR), SVM), NB, Árbol de Decisión (del inglés *Decision Tree* o DT), Red Neuronal Convolucional (del inglés *Convolutional Neural Network* o CNN) y LSTM; y traducción automática con Traducción Automática Neuronal (del inglés *Neural Machine Translation* o NMT) y Traducción Automática Estadística (del inglés *Statistical Machine Translation* o SMT).

2.6.4 Similitud Semántica en Textos Cortos

2.6.4.1 *A survey on the techniques, applications, and performance of short text semantic similarity* (Han et al., 2021)

Este estudio compara métodos para medir similitud semántica: Codificación binaria de categorías en vectores dispersos (del inglés *One-hot encoding* o ONE-HOT), Modelo de espacio vectorial para representación de documentos (del inglés *Vector Space Model* o VSM), *Word-2Vec*, *WordNet* y aprendizaje profundo con Memoria a Corto y Largo Plazo Siamés (del inglés *Siamese Long Short-Term Memory* o SiaLSTM), Memoria a Corto y Largo Plazo Bidireccional Siamés (del inglés *Siamese Bidirectional Long Short-Term Memory* o SiaBiLSTM), Memoria a Corto y Largo Plazo Bidireccional Siamés con Mecanismo de Atención (del inglés *Attention-based Siamese Bidirectional Long Short-Term Memory* o ATTSiaBiLSTM) y BERT. Se evaluó con Error Cuadrático Medio, Precisión y $F1$-*score*. BERT obtuvo un Error Cuadrático Medio (del inglés *Mean Squared Error* o MSE) de 0.137, Exactitud de 0.923 y $F1$-*score* de 0.899, superando a ATTSiaBiLSTM, MSE) 0.175, ACC 0.828, $F1$-*score* 0.747) y a todos los métodos previos.

2.6.4.2 *FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference* (Martín et al., 2022)

El estudio evalúa la extracción de palabras clave en Twitter comparando FactTeR-CheckKey, KeyBERT y Extracción automática de palabras clave mediante análisis de frecuencia y co-ocurrencia (del inglés *Rapid Automatic Keyword Extraction* o RAKE), utilizando precisión, exhaustividad y $F1$-*score*. FactTeR-CheckKey con filtrado Etiquetado de categorías gramaticales de las palabras en un texto (del inglés *Part-of-Speech Tagging* o POS) y tamaño grande alcanzó 0.6888 de precisión, 0.7917 de *recall* y 0.7149 de $F1$-*score*, superando a KeyBERT (0.6483, 0.6242, 0.6107) y RAKE (0.5707, 0.8590, 0.6653), mostrando mejor equilibrio entre precisión y cobertura.

2.6.4.3 *How can we know what language models know?* (Jiang et al., 2020)

El artículo propone descubrir automáticamente mejores *Prompts* mediante minería y parafraseo, combinando respuestas con métodos de conjunto. En BERT-*large*, la combinación de minería automática y parafraseo manual de *Prompts* para generar variantes efectivas (del inglés *Combination of automatic Prompt Mining and manual paraphrasing to generate effective variants* o Mine+Man) obtuvo 39.4 % en Mejor opción seleccionada según la métrica principal (del inglés *Top-1 selection according to main metric* o Top1) y 43.9 % en Selección óptima de *Prompts* basada en combinación o evaluación (del inglés *Optimal Prompt selection based on combination or evaluation* o OPTI) bajo precisión micro promediada, y 28.1 % en Top1 y 30.7 % en OPTI bajo precisión macro promediada. Los valores corresponden a métricas de precisión expresadas en porcentaje.

2.6.4.4 *Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives* (Amur et al., 2023)

El artículo revisa desafíos y mejoras en sistemas de preguntas cortas, diseñados para responder preguntas breves con precisión usando modelos de lenguaje (del inglés *Short Question Answering Systems, designed to answer brief questions accurately using language models* o SQAS) usando modelos como BERT. Se aplicaron atención, optimización bayesiana, búsqueda en cuadrícula y codificación universal para mejorar contexto y precisión. BERT alcanzó 98.00 % de precisión y 91.50 % de $F1$-*score* en detección de sarcasmo; ELMO obtuvo 80.00 % precisión en similitud y 71.30 % *Pearson* r en similitud Similitud de Pregunta a Pregunta (del inglés *Question-to-Question Similarity* o Q2Q) árabe.

2.6.5 Resumen de la Exploración del Estado del Arte

A continuación, se presenta una síntesis estructurada de los trabajos analizados durante la revisión del estado del arte. La Tabla 2.1 organiza de manera comparativa los estudios seleccionados, considerando sus modelos, técnicas, métricas y principales hallazgos. Este resumen permite visualizar de forma integrada las tendencias metodológicas y los resultados más relevantes en investigaciones recientes relacionadas con *Prompt learning*, LLMs y técnicas de evaluación.

Tabla 2.1. Comparativa de revisión de la literatura.

Id	Título	Modelo(s) o Framework(s)	Técnica(s)	Métrica(s)	Resultados
1	*Instance-aware Prompt Learning for Language Understanding and Generation* (Jin et al., 2023)	GPT2-*base*, GPT2-*large*	IPL, FT, PT	*BERTScore*, BLEU, NIST, ROUGE-L, CIDEr, METEOR, TER	IPL supera consistentemente a FT y PT. **GPT2-*base*:** *BERTScore* 95.43 %, BLEU 69.82 %, NIST 8.82, ROUGE-L 71.65 %, CIDEr 2.49, METEOR 72.23, TER 48.50 **GPT2-*large*:** *BERTScore* 95.47 %, BLEU 68.53 %, NIST 8.68, ROUGE-L 71.20 %, CIDEr 2.51, METEOR 50.55, TER 46.17

Tabla 2.1. Comparativa de revisión de la literatura (*continuación*).

Id	Referencia	Modelo(s) o Framework(s)	Técnica(s)	Métrica(s)	Resultados
2	*Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing* (Liu et al., 2023)	L2R, ERNIE, GPT-1, GPT-2, RoBERTa, mBERT, BERT, XLM-R, PEGASUS, ELMo, BART, GLM, CTRL, GPT-3, PanGu-a, T5, ALBERT, CPM-2, Trans, GPT-like, GPT, Conv, ERNIE-B3	Aprendizaje basado en *Prompts*, pre-entrenamiento, adaptación con pocos ejemplos	Número de tareas por año	Entre 2018-2021, aumentaron investigaciones, con 29 estudios en 2021. Las tareas incluyen: - Clasificación de texto: 22 - Sondeo factual: 15 - Generación condicional: 8
3	*Recent advances in natural language processing via large pre-trained language models: A survey* (Min et al., 2023)	BERT, RoBERTa*large*	Tareas proxy estilo preguntas y respuestas con plantillas de contexto “C”, *Prompt* “Q” y respuesta “A”	Técnica de evaluación basada en precisión de respuestas generadas en clasificación y extracción	Usado en clasificación de emociones y extracción de argumentos, donde las respuestas generadas se evalúan según su precisión contextual.
4	*How can we know what language models know?* (Jiang et al., 2020)	BERT-*large*	Descubrimiento automático de mejores *Prompts* mediante minería y parafraseo, combinación con métodos de conjunto	Precisión micro promediada, Precisión macro promediada (porcentaje)	Método combinado Mine+Man alcanzó 39.4 % (Top1) y 43.9 % (OPTI) en precisión micro promediada, y 28.1 % (Top1) y 30.7 % (OPTI) en precisión macro promediada.
5	*Measuring and Improving Consistency in Pretrained Language Models* (Elazar et al., 2021)	BERT, RoBERTa, ALBERT, BERT-*large*-wwm	Consistencia en extracción de conocimiento mediante PARAREL (328 patrones y 38 relaciones semánticas)	*Accuracy*, *Consistency*, *Consistent-Acc*	*Accuracy* y *Consistency* fueron 48.7 % ± 25.0 y 60.9 % ± 24.2, respectivamente.

Tabla 2.1. Comparativa de revisión de la literatura (*continuación*).

Id	Referencia	Modelo(s) o Framework(s)	Técnica(s)	Métrica(s)	Resultados
6	*Improving patient self-description in Chinese online consultation using contextual Prompts* (X. Li et al., 2022)	BERT, BoW	Modelos sin *Prompt*s, *Prompts* aprendidos, *Prompts* basados en certeza e incertidumbre	*accuracy*, *macro* $F1$-*score*, *macro* AUC, *macro* MCC, *entropy*	Modelos con *Prompts* aprendidos mejoran *precision* (0.33–0.70), AUC (0.80–0.92) y MCC (hasta 0.56) frente a heurísticas. Estrategias basadas en certeza e incertidumbre no mostraron mejora. Evaluación en tres corpus clínicos.
7	*Personalized Prompt Learning for Explainable Recommendation* (L. Li et al., 2023)	GPT-2, BERT, GPT-3	PEPLER-D, *Sequential Tuning*, *Prompt*+LM, FT, *Fixed*-M PT	Unidad de medida: técnica (evaluación mediante BLEU-4 y *learning rate*)	PEPLER-D mejora la explicabilidad en recomendación, superando a BERT y GPT-3 en BLEU-4: 0.96 (*Amazon*), 0.9 (*TripAdvisor*), 0.69 (*Yelp*). *Learning rate* entre 10^{-5} y 10^{-3}.
8	*NIR-Prompt: A Multi-task Generalized Neural Information Retrieval Training Framework* (Xu et al., 2023)	BERT-*base* (109M)	NIR-*Prompt*, *tokens* fijos de indicación durante entrenamiento	Técnica (evaluación mediante *Precision*, *Accuracy*, $F1$-*score*)	Mejor rendimiento con $k = 11$. Paráfrasis: *Accuracy* 79.8 %, $F1$-*score* 83.8 %. Inferencia de Lenguaje Natural: *Accuracy* 77.5 %, $F1$-*score* 84.5 %.
9	*Comprehensive Empirical Study of Bias Mitigation Methods for LM Classifiers* (Chen et al., 2023)	LMs	Mitigación de sesgo en LM, evaluación de 17 métodos	Metodología cuantitativa (11 métricas de rendimiento, 4 de equidad, 20 de compensación)	Reducción de rendimiento entre 42 % y 66 %. Mejora en equidad de 24 % a 59 %. Compromisos en 25 % de escenarios. Ningún método es óptimo universalmente; uno destacó en 30 % de casos.
10	*From Natural Language Processing to Neural Databases* (Thorne et al., 2021)	Modelos de transformadores de NLP, T5, DPR	Sistema *NeuralDB* para consultas en lenguaje natural sin esquema	Porcentaje de coincidencia exacta	*NeuralDB* logró 79.45 % en consultas *Count* y 100 % en Min/Max *Sets*, superando a TF-IDF + T5 (31.06 %) y DPR + T5 (38.07 %). Métrica: precisión exacta.

Tabla 2.1. Comparativa de revisión de la literatura (*continuación*).

Id	Referencia	Modelo(s) o Framework(s)	Técnica(s)	Métrica(s)	Resultados
11	*Knowledge enhanced Prompt tuning for Stance Detection* (Huang et al., 2023)	Métodos tradicionales de NLP, aprendizaje profundo, datasets multilingües (inglés, chino, español)	Detección automática de postura con *Prompt*'s tuning y aprendizaje automático	$F1$-*score*, *Precision*, *Recall*, Exactitud	Evaluación cuantitativa en datasets multilingües (inglés, chino, español) con $F1$-*score*, *Precision*, *Recall* y Exactitud. Técnicas tradicionales y aprendizaje profundo mejoran detección de postura.
12	*A survey on the techniques, applications, and performance of short text semantic similarity* (Han et al., 2021)	*One-hot*, VSM, Word2Vec, WordNet, SiaLSTM, SiaBiLSTM, ATTSiaBiLSTM, BERT	Métodos para medir similitud semántica en texto corto	Error Cuadrático Medio, *Precision*, $F1$-*Score*	Medición cuantitativa de similitud semántica en texto corto usando MSE, *Precision* y $F1$-textitscore. BERT supera métodos previos con MSE 0.137, *Accuracy* 0.923 y $F1$-*score* 0.899; ATTSiaBiLSTM logró MSE 0.175, *Accuracy* 0.828 y $F1$-*score* 0.747.
13	*OCTIS: Comparing and optimizing topic models is simple!* (Terragni et al., 2021)	Gensim, MALLET, OCTIS	Optimización bayesiana para ajuste de hiperparámetros en modelos de temas	Coherencia, diversidad, significancia, clasificación	Optimización bayesiana de hiperparámetros en modelos de temas con OCTIS, que compara modelos clásicos y neuronales mediante coherencia, diversidad y clasificación. Destaca por interfaz web, sin un modelo superior definitivo.
14	*The evolution of topic modeling* (Churchill & Singh, 2022)	LDA, NMF, VAETM	Aprendizaje automático y redes neuronales para mejora de coherencia y diversidad temática	Coherencia, diversidad, desempeño en identificación de tópicos relevantes	Estudio de la evolución de modelos de tópicos (LDA, NMF, VAETM) usando aprendizaje automático y redes neuronales para mejorar coherencia, diversidad y desempeño en identificación temática. Factores clave: tipo de datos, ruido y tareas posteriores.

Tabla 2.1. Comparativa de revisión de la literatura (*continuación*).

Id	Referencia	Modelo(s) o Framework(s)	Técnica(s)	Métrica(s)	Resultados
15	*Natural language processing for Nepali text: a review* (Shahi & Sitaula, 2022)	NB, SVM, ANN, LSTM, k-means, *Fuzzy Set*, LR, DT, CNN, NMT, SMT	Aprendizaje automático, transferencia de conocimiento, creación de recursos lingüísticos, análisis de sentimientos, traducción automática	Precisión, clasificación, agrupación, análisis de sentimientos, traducción automática	Revisión de técnicas de NLP para texto nepalí (NB, SVM, ANN, LSTM, CNN, NMT, entre otras). Se enfoca en aprendizaje automático, transferencia de conocimiento, recursos lingüísticos, análisis de sentimientos y traducción automática. Métricas: precisión y clasificación. Destaca estrategias de transferencia y grandes datasets anotados para mejorar desempeño.
16	*FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference* (Martín et al., 2022)	FactTeR-CheckKey: sistema para extracción de palabras clave y verificación de hechos, KeyBERT: técnica basada en BERT para extracción de palabras clave, RAKE	Extracción de palabras clave, similitud semántica, inferencia de lenguaje natural	*Precision*, exhaustividad, $F1$-*score*	FacTeR-Check: verificación semi-automática de hechos mediante extracción de palabras clave, similitud semántica e inferencia natural. Métodos: FactTeR-CheckKey, KeyBERT, RAKE. Métricas: *precision*, *recall*, $F1$-*score*. FactTeR-CheckKey logró *precision* 0.6888, *recall* 0.7917 y $F1$-*score* 0.7149, superando a otros métodos en Twitter.

Tabla 2.1. Comparativa de revisión de la literatura (*continuación*).

Id	Referencia	Modelo(s) o Framework(s)	Técnica(s)	Métrica(s)	Resultados
17	*How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN* (McCoy et al., 2023)	GPT-2 (cuatro tamaños)	Medición de novedad lingüística, análisis de duplicación por tamaño de n-grama	Análisis cuantitativo de duplicación y novedad	Análisis de novedad lingüística en generación de texto con GPT-2 (4 tamaños) usando RAVEN. Estudio de duplicación y novedad por tamaño de n-grama. GPT-2 produce estructuras nuevas pero replica fragmentos, usando generalización composicional y analógica con algunas fallas semánticas. Tamaño del modelo no determina novedad.
18	*Impact of word embedding models on text analytics in deep learning environment: a review* (Asudani et al., 2023)	LSTM, Word2Vec, CNN, BERT, fastText	Incrustaciones de palabras específicas de dominio, aprendizaje profundo (LSTM, CNN)	*Precision*, $F1$-*score*	Revisión de incrustaciones de palabras y aprendizaje profundo (LSTM, CNN, BERT, fastText) en análisis de texto. Evaluación cuantitativa con *precision* y $F1$-*score*. LSTM + Word2Vec 99.55 % *precision*; CNN + Word2Vec chino 96.60 %; BERT 95.00 %; fastText 91.00 % $F1$-*score*. Incrustaciones específicas y memoria a largo/corto plazo mejoran rendimiento.

Tabla 2.1. Comparativa de revisión de la literatura (*continuación*).

Id	Referencia	Modelo(s) o Framework(s)	Técnica(s)	Métrica(s)	Resultados
19	*Topic modeling algorithms and applications: A survey* (Abdelrazek et al., 2023)	LDA, NMF, VAETM, redes neuronales	Aprendizaje automático, modelado de temas, análisis de coherencia y diversidad temática	Coherencia, diversidad temática	Revisión de algoritmos de modelado de temas (LDA, NMF, VAETM, redes neuronales) con énfasis en aprendizaje automático, coherencia y diversidad temática. Evaluación cuantitativa de coherencia y diversidad. Factores clave: tamaño, ruido, tipo de datos y tareas posteriores. Adaptar modelos al contexto mejora identificación de tópicos.
20	*Short-Text Semantic Similarity: Techniques, Challenges and Future Perspectives* (Amur et al., 2023)	BERT, ELMO	Atención, optimización bayesiana, búsqueda en cuadrícula, codificación universal	*Precision*, $F1$-*score*, Pearson r	BERT y ELMO en similitud semántica de texto corto. Evaluación cuantitativa con *precision*, $F1$-*score* y Pearson r. BERT: 98.0 % *precision* y 91.5 % $F1$-*score* en detección de sarcasmo; ELMO: 80.0 % *precision* y 71.3 Pearson r en similitud Q2Q árabe. Atención y optimización bayesiana mejoran contexto y *precision*.
21	*Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking* (Y.-C. Zhou et al., 2022)	Aprendizaje profundo, NLP, análisis sintáctico, etiquetado semántico	Interpretación automatizada de reglas, análisis sintáctico, etiquetado semántico	*Precision*, *recall*, $F1$-*score*	*Precision*, *recall* y $F1$-*score* promedio alrededor de 86 % en interpretación automatizada de reglas para *compliance*. Evaluación sobre 4,575 etiquetas. Método mejora comprensión de reglas complejas.

Tabla 2.1. Comparativa de revisión de la literatura (*continuación*).

Id	Referencia	Modelo(s) o Framework(s)	Técnica(s)	Métrica(s)	Resultados
22	*Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering* (Velásquez-Henao et al., 2023b)	Modelos de lenguaje de inteligencia artificial	Metodología GPEI, diseño iterativo de *Prompts* precisos y contextuales, inteligencia artificial explicable, evaluación continua de respuestas	Metodología de optimización y evaluación continua	Metodología cualitativa que promueve *Prompts* transparentes y mejora la comunicación con LLMs mediante diseño iterativo y evaluación continua.
23	*Interactive and visual Prompt engineering for ad-hoc task adaptation with large language models* (Strobelt et al., 2022)	LLMs	Sistema interactivo PromptIDE para personalización y refinamiento visual de *Prompts*, exploración con visualizaciones	Resultados cuantitativos de evaluación en clasificación y refinamiento de *Prompts*	Evaluación cuantitativa con 84 respuestas correctas y 16 incorrectas. Clasificación precisa en política, deportes, negocios y ciencia. PromptIDE mejora personalización e interacción sin necesidad de experiencia previa.
24	*Jointly improving parsing and perception for natural language commands through human-robot dialog* (Thomason et al., 2020)	Agentes robóticos, NLP, aprendizaje interactivo humano-robot	Metodología de aprendizaje interactivo integrando percepción física y comprensión del lenguaje natural	Probabilidades numéricas para evaluación de atributos en objetos	Evaluación con probabilidades numéricas (0.04 a 0.56) para atributos físicos. Mejora coherencia y precisión en diálogo humano-robot, integrando percepción y NLP.

2.6.6 Discusión

Los principales hallazgos de los trabajos revisados destacan la relevancia de los LLMs basados en transformadores (como BERT y GPT) en el ámbito del NLP, especialmente en tareas como la generación de texto y la extracción de información. En particular, los enfoques de preentrenamiento y ajuste fino, el aprendizaje basado en *Prompts* y la reformulación de tareas han resultado fundamentales para mejorar el rendimiento de estos modelos (Min et al., 2023). Además, la literatura señala que la efectividad de los *Prompts* depende en gran medida de su formulación, pudiendo optimizarse mediante técnicas como minería de datos, paráfrasis y combinación de respuestas de múltiples *Prompts* (Jiang et al., 2020).

En este sentido, enfoques como CoOP han mostrado que la automatización en la creación de *Prompts* puede superar las limitaciones de las versiones manuales, incrementando la robustez y favoreciendo la transferencia a tareas de distintos dominios (K. Zhou et al., 2022). De

manera complementaria, técnicas recientes en similitud semántica basadas en redes neuronales profundas han demostrado mejorar la precisión en clasificación, resumen automático y traducción, superando a los métodos tradicionales y consolidándose como alternativas más efectivas para procesar información semántica (Chandrasekaran & Mago, 2021). Estos avances refuerzan la necesidad de explorar enfoques que combinen refinamiento automático de entradas con estrategias de evaluación de calidad.

Asimismo, (Han et al., 2021) evidencia que el uso de modelos como BERT mejora significativamente la precisión en la medición de similitud semántica aplicada a tareas como clasificación y análisis de sentimientos. Esto confirma la utilidad de los LLMs en la extracción de información relevante y su efectividad en contextos como redes sociales y textos breves. Tales hallazgos son clave para esta investigación, pues refuerzan la importancia de diseñar mecanismos que permitan evaluar y optimizar *Prompts* en dominios específicos.

De manera más concreta, (Jin et al., 2023), en el artículo "*Instance-aware Prompt Learning for Language Understanding and Generation*", implementaron un enfoque de aprendizaje de *Prompts* basado en instancias, logrando mejoras significativas en el rendimiento de modelos GPT2-*base* y GPT2-*large* en tareas de generación de lenguaje natural. El método *Instance-aware Prompt Learning* superó a enfoques de FT y PT en métricas como BLEU, NIST, ROUGE-L y BERTScore. Por ejemplo, en generación de texto a partir de tablas Extremo a Extremo (del inglés *End-to-End* o E2E), obtuvo un puntaje BLEU de 69.82 en GPT2-*base* y 68.53 en GPT2-*large*, mientras que los modelos de referencia alcanzaron resultados menores. En la tarea de resumen SamSum, logró un ROUGE-L de 42.0 en GPT2-*base* y 45.0 en GPT2-*large*, evidenciando un rendimiento superior.

Estos resultados se asemejan a la propuesta de esta tesis, ya que también buscan optimizar la formulación de los *Prompts* para mejorar la calidad de las respuestas generadas por los LLMs. No obstante, mientras que los trabajos revisados se enfocan principalmente en inglés y en contextos restringidos, la presente investigación aborda un vacío poco explorado: el preprocesamiento de *Prompts* redactados en español culto por usuarios no expertos. Con ello, se pretende contribuir a la creación de criterios de evaluación y estrategias de refinamiento que favorezcan interacciones más claras, consistentes y efectivas entre los usuarios y los modelos de lenguaje.

Capítulo 3

Diseño e Implementación de la Propuesta de Solución

En el ámbito de los LLMs, la calidad de los *Prompts* desempeña un papel fundamental en la interacción con estos modelos. Este capítulo desarrolla la propuesta metodológica desde tres niveles: conceptual, técnico y gráfico.

El nivel conceptual describe las etapas del proceso metodológico que guían el diseño de la solución; el nivel técnico detalla las herramientas, procedimientos y tecnologías utilizadas para su implementación; y el nivel gráfico presenta la representación visual de la metodología, la cual se incluye en la Sección 3.2.

3.1 Planteamiento de la Solución

La propuesta de esta investigación busca resolver la problemática identificada en torno a la elaboración de *Prompts* efectivos en español culto para su uso con LLMs. La variabilidad en la calidad de las entradas generadas por usuarios no expertos limita la precisión, coherencia y utilidad de las respuestas producidas por estos modelos.

Con base en ello, se plantea el diseño de una metodología que combine principios de NLP y técnicas de Ingeniería de *Prompts*, orientada a mejorar la estructura, claridad y efectividad de las instrucciones en lenguaje natural. Dicha metodología permitirá guiar al usuario en la creación de *Prompts* comprensibles y consistentes, contribuyendo al fortalecimiento de la interacción humano–modelo.

La metodología propuesta se organiza en seis etapas, que abarcan desde la recolección y preparación de datos hasta la validación de los resultados experimentales. Una representación gráfica general de este proceso se muestra en la Figura 3.1, ubicada en la Sección 3.2.

3.2 Metodología de Trabajo

La metodología propuesta constituye el eje central de esta investigación, ya que integra de manera estructurada las etapas necesarias para la recolección, análisis, mejora y validación de los *Prompts* elaborados por usuarios no expertos.

La Figura 3.1 presenta la visión general del proceso metodológico, compuesto por seis etapas interrelacionadas. Cada una de ellas define un conjunto de actividades que permiten avanzar desde la obtención inicial de datos hasta la evaluación final de la hipótesis.

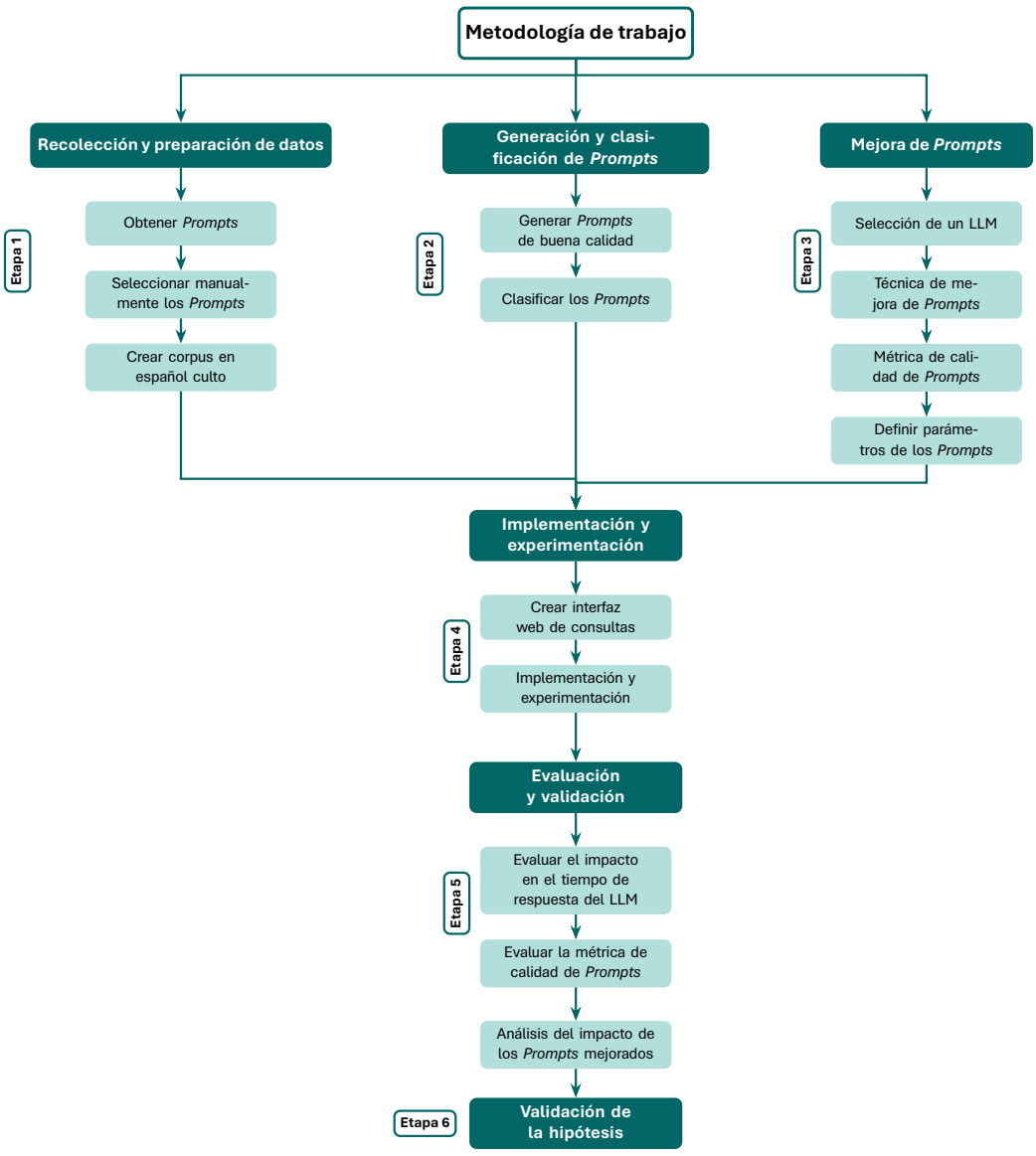

Figura 3.1: Diagrama general de la metodología propuesta.

En el diagrama se observa cómo las seis etapas —*Recolección y preparación de datos*, *Generación y clasificación de Prompts*, Mejora de *Prompts*, Implementación y experimentación, Evaluación y validación, y finalmente, Validación de la hipótesis— conforman un flujo metodológico iterativo.

Este enfoque garantiza que la propuesta no sólo evalúe la calidad de los *Prompts*, sino que también permita su mejora continua mediante retroalimentación basada en métricas de desempeño y análisis de resultados experimentales.

3.2.1 Etapa 1: Recolección y Preparación de Datos

Esta etapa tiene como objetivo construir el conjunto inicial de *Prompts* y textos de referencia que servirán como base para la generación y mejora posterior.

- **Obtención de *Prompts*:** Se recopiló un total de 1,796 *Prompts* en español culto desde el foro */preguntaleareddit* de la plataforma Reddit. En una primera extracción se obtuvieron 883 *Prompts*, que se utilizaron para la etapa inicial de diseño de la métrica de calidad de *Prompts*. Posteriormente, se incorporaron 913 *Prompts* adicionales, conformando el conjunto completo, que se empleó para análisis estadísticos destinados a establecer umbrales de calidad y longitud de texto del *Prompt*.

  Por otro lado, se utilizó el *dataset Multilingual BioASQ-6B*, especializado en temas de biología, para aprovechar su estructura tipo pregunta-respuesta redactada por expertos humanos. Este *dataset* permitió realizar un análisis estadístico orientado a determinar el tamaño óptimo de las respuestas generadas por un LLM. En total, se emplearon 2,251 *Prompts*, cada uno acompañado de su respectiva respuesta, lo que proporcionó una base sólida para evaluar el comportamiento del modelo en tareas de generación controlada de texto.

- **Selección manual de *Prompts*:** Se filtraron aquellos con calidad lingüística baja, con el fin de crear ejemplos contrastivos para la mejora posterior.

- **Construcción del corpus en español culto:** Se elaboró un corpus basado en textos de referencia, particularmente de la revista Arqueología Mexicana (Ed. esp. 115) (Arqueología Mexicana, 2024), que sirve como material para entrenamiento y comparación.

Esta fase asegura que los datos iniciales sean representativos, consistentes y pertinentes para la evaluación de la metodología.

3.2.2 Etapa 2: Generación y Clasificación de *Prompts*

Esta etapa tiene como propósito ampliar y estructurar el conjunto de datos obtenido en la Etapa 1 mediante la generación de ejemplos representativos y su clasificación manual de acuerdo con criterios lingüísticos. El proceso se organiza en dos fases complementarias: generación de ejemplos buenos y malos, y clasificación manual de los *Prompts*.

Generación de ejemplos buenos y malos

Los 1,796 *Prompts* obtenidos desde el foro */preguntaleareddit* de la plataforma Reddit se seleccionaron inicialmente por su baja calidad lingüística, caracterizada por errores ortográficos, gramaticales y construcciones informales. A partir de este material, se generaron versiones mejoradas con la asistencia del LLM GPT-4o, con el objetivo de transformar los textos originales en *Prompts* tipo pregunta, redactados con claridad, coherencia y corrección gramatical.

De esta manera, el conjunto final está conformado por 1,796 *Prompts* originales y 1,796 *Prompts* reformulados tipo pregunta, sumando un total de 3,592 *Prompts*. Los ejemplos "malos" representan errores comunes en la escritura informal, mientras que los ejemplos "buenos" muestran formulaciones claras y estructuradas que sirven como referencia de calidad lingüística.

Este proceso de generación controlada permite construir un conjunto contrastivo de *Prompts*, que posteriormente se utiliza en la Etapa 3 para el diseño de técnicas de mejora y validación del enfoque propuesto.

Clasificación manual

Una vez generadas las versiones contrastivas, los *Prompts* se clasifican de forma manual en dos categorías principales:

- **Tipo 0:** *Prompts* originales con errores ortográficos, gramaticales o de estilo informal.
- **Tipo 1:** *Prompts* corregidos o reformulados mediante GPT-4o, con una estructura interrogativa clara y adecuada al registro del español culto.

Este proceso garantiza que el conjunto de datos refleje contrastes reales en la calidad de redacción, lo cual resulta esencial para las etapas posteriores de mejora y evaluación.

3.2.3 Etapa 3: Mejora de *Prompts*

La tercera etapa del proceso se centra en la mejora de *Prompts*, buscando optimizar su claridad, estructura y efectividad para su interacción con LLMs. Esta etapa abarca tres compo-

nentes principales: selección del LLM, técnica de mejora basada en NLP y definición de parámetros de calidad.

3.2.3.1 Selección del LLM

Para la implementación de la mejora automática de *Prompts*, se evaluaron distintos LLMs, considerando su capacidad de análisis lingüístico, generación de texto coherente y adecuación al español. Se realizaron pruebas exploratorias con LLaMA, Mistral y GPT-2:

- **LLaMA**: desarrollado por Meta, orientado a tareas de NLP, pero requirió altos recursos computacionales y almacenamiento local.
- **Mistral**: modelo de código abierto optimizado para generación de texto, también demandante en recursos.
- **GPT-2**: mostró limitaciones en comprensión de contexto en español.

Finalmente, se seleccionó **GPT-4o** de OpenAI por su accesibilidad, robustez en generación de texto en español, escalabilidad en la nube y capacidad para análisis lingüístico profundo. Además, su integración mediante API permitió desarrollar la interfaz de mejora automática de forma eficiente.

3.2.3.2 Técnica de Mejora Basada en NLP

La técnica de mejora propuesta consiste en reformular *Prompts* mal estructurados hacia un formato tipo pregunta adecuado para la interacción con LLMs. El proceso busca corregir errores lingüísticos y optimizar la estructura, la semántica y el contexto, de modo que el *Prompt* resultante sea más claro y efectivo para la generación de texto.

3.2.3.3 Definición de Parámetros

Los parámetros de evaluación y mejora de *Prompts* se definieron considerando tres dimensiones principales:

- **Longitud**: se busca un equilibrio entre dar suficiente información y ser claro al escribir, para evitar *Prompts* confusos o demasiado largos.
- **Claridad**: se evalúa la ausencia de errores lingüísticos (ortografía, sintaxis, puntuación, gramática) y la coherencia del mensaje.
- **Contexto**: se analiza la relevancia semántica y la adecuación al dominio, garantizando que el *Prompt* sea comprensible y específico para el LLM.

3.2.4 Etapa 4: Implementación y Experimentación

En esta etapa se desarrolla el prototipo funcional del sistema, orientado a integrar la técnica de mejora de *Prompts* dentro de una interfaz web. La implementación se centra en tres aspectos principales: la construcción de la interfaz, la ejecución de pruebas controladas y la integración con el LLM seleccionado.

3.2.4.1 Construcción de la Interfaz Web

Se desarrolla una interfaz web con el propósito de demostrar el funcionamiento básico del sistema de mejora de *Prompts*. Esta interfaz permite ingresar un *Prompt*, enviarlo al LLM y recibir una versión mejorada en tiempo real.
La estructura se diseña con un enfoque modular, incluyendo:

- Un campo de entrada para capturar el texto del *Prompt*.
- Un botón de procesamiento que envía la solicitud al servidor.
- Un área de salida donde se muestran las respuestas generadas por el modelo.

3.2.4.2 Ejecución de Pruebas Controladas

Se realizan pruebas controladas para asegurar el correcto funcionamiento del sistema en condiciones estables. Estas pruebas verifican:

- La comunicación entre la interfaz web y la API del LLM.
- La correcta recepción, envío y procesamiento de los *Prompts*.
- El tiempo de respuesta.

3.2.4.3 Integración con el LLM Seleccionado

Para la integración del sistema se utiliza **GPT-4o** de OpenAI, elegido por su alto rendimiento en tareas de generación y análisis lingüístico en español. La conexión se realiza mediante su API en la nube, lo que permite un acceso remoto, estable y escalable.
También se realizan pruebas preliminares con modelos alternativos como LLaMA, Mistral y GPT-2. Sin embargo, estos modelos presentan limitaciones técnicas: mayores requerimientos de hardware y menor comprensión del contexto en español. Por ello, GPT-4o se selecciona para la versión final de la interfaz web.
Con esta implementación se establecen las bases técnicas necesarias para iniciar, en la siguiente sección, la experimentación práctica y la validación de la técnica de mejora propuesta.

3.2.5 Etapa 5: Evaluación y Validación

En esta etapa se evalúa la efectividad de la técnica de mejora de *Prompts* mediante una métrica de calidad que cuantifica objetivamente la mejora en los *Prompts*. La evaluación considera varios aspectos: la calidad lingüística, el impacto en el tiempo de respuesta del LLM, la longitud óptima de los *Prompts* y de las respuestas generadas, así como la comparación entre los *Prompts* originales y los mejorados.

3.2.5.1 Métricas de Calidad de los *Prompts*

Se define una métrica de calidad capaz de medir múltiples características lingüísticas de los *Prompts*, incluyendo puntuación, ortografía, sintaxis, gramática, fluidez, léxico, registro, pragmatismo, contexto, incoherencia y semántica. Esta métrica asigna un valor cuantitativo a cada *Prompt*, el cual refleja su calidad relativa, pero no determina por sí solo si un *Prompt* es bueno o malo. Para clasificar los *Prompts* según su calidad, se realiza un análisis estadístico adicional que establece el umbral correspondiente. Además, se definen umbrales óptimos de longitud para los *Prompts* y para las respuestas generadas por el LLM, garantizando claridad y precisión en la interacción.

3.2.5.2 Evaluación del Impacto en el Tiempo de Respuesta

Para complementar la evaluación de calidad, se mide el tiempo que el LLM tarda en generar respuestas ante *Prompts* originales y mejorados. Se determina si la técnica de mejora contribuye a una reducción en el tiempo de respuesta, indicando un mejor entendimiento del *Prompt* por parte del modelo.

3.2.5.3 Comparación entre *Prompts* Originales y Mejorados

La comparación entre *Prompts* originales y mejorados se realiza considerando la calidad lingüística, el cumplimiento de los umbrales de longitud, el valor de calidad respecto al umbral definido y el tiempo de respuesta. Se aplican pruebas estadísticas para identificar diferencias significativas entre ambos grupos. Asimismo, se emplean visualizaciones gráficas, como boxplots, para observar la distribución de los tiempos de respuesta y la dispersión de los valores de calidad.

Los criterios para considerar un *Prompt* mejorado como exitoso incluyen:

- Presentar una estructura clara y coherente.
- Cumplir con los umbrales establecidos de calidad y longitud.
- Mostrar reducción en el tiempo de respuesta del LLM.
- Registrar mejoras perceptibles en todas las características lingüísticas evaluadas respecto al *Prompt* original.

Con esta etapa se establece evidencia objetiva sobre la efectividad de la técnica de mejora, sentando las bases para la validación de la hipótesis de investigación.

3.2.6 Etapa 6: Validación de la Hipótesis

En esta etapa se valida la hipótesis central de investigación: *Aplicar Ingeniería de Prompts mejora la calidad de los Prompts generados por usuarios no expertos*. La validación se realiza mediante un análisis cuantitativo y cualitativo de los resultados obtenidos en la etapa anterior.

3.2.6.1 Análisis Cuantitativo

Se comparan los *Prompts* originales y los mejorados utilizando la métrica de calidad definida previamente, considerando todas las características lingüísticas evaluadas, así como los umbrales óptimos de longitud de los *Prompts* y de las respuestas generadas por el LLM. Se incluyen también los tiempos de respuesta del LLM, midiendo si los *Prompts* mejorados permiten una interacción más rápida y eficiente.

Para determinar si la mejora es significativa, se aplican pruebas estadísticas, incluyendo pruebas de normalidad, pruebas no paramétricas (como Wilcoxon) y análisis de varianza cuando corresponde. Estas pruebas cuantitativas permiten establecer objetivamente si los *Prompts* mejorados superan el umbral de calidad definido y cumplen con los criterios de efectividad establecidos.

3.2.6.2 Análisis Cualitativo

Se realiza un análisis cualitativo de los *Prompts* mejorados, evaluando aspectos como claridad, coherencia, adecuación contextual y corrección lingüística. Este análisis complementa los resultados cuantitativos, comparando las mejoras obtenidas con métricas tradicionales de calidad de texto, como coherencia, relevancia y diversidad, e identificando cambios en la estructura y en la formulación de las preguntas que facilitan la comprensión y la generación de respuestas precisas por parte del LLM.

3.2.6.3 Confirmación de la Hipótesis

La hipótesis se considera validada si los *Prompts* mejorados muestran una calidad significativamente mayor que los *Prompts* originales, superando los umbrales definidos para cada métrica, y si además se observa una reducción consistente en el tiempo de respuesta del modelo. La combinación del análisis cuantitativo y cualitativo proporciona evidencia sólida sobre la efectividad de la técnica de mejora de *Prompts* aplicada.

3.3 Modelo Conceptual de la Solución

El modelo conceptual ilustra de manera integral el flujo de datos y procesos que conforman la metodología propuesta para la mejora de *Prompts*. Representa la secuencia de interacción entre los módulos principales del sistema —desde la captura inicial del *Prompt* hasta la obtención de la respuesta optimizada—, permitiendo comprender la dinámica interna del proceso y su correspondencia con las etapas metodológicas descritas.

Como se observa en la Figura 3.2, el modelo conceptual describe cómo los *Prompts* sin procesar son introducidos por el usuario a través de una interfaz web, evaluados mediante un módulo de medición de calidad y posteriormente optimizados por un componente de aprendizaje automático (LLM con ajuste fino). Tras la mejora, los nuevos *Prompts* son revaluados y, si cumplen con los criterios de calidad establecidos, se envían al modelo de lenguaje grande (LLM) para la generación de la respuesta final.

De manera estructurada, el flujo se organiza en tres niveles:

- **Entradas:** *Prompts* sin procesar proporcionados por el usuario a través de la interfaz del sistema web. Estos constituyen el punto de partida del proceso de mejora y evaluación de calidad.
- **Procesamiento:** ejecución de la técnica de mejora de *Prompts* utilizando el modelo GPT-4o como sistema base. En esta etapa se evalúa la efectividad de los *Prompts*, se aplica un proceso de optimización iterativa y se analizan variables de desempeño del modelo, como la longitud del *Prompt*, la extensión de la respuesta generada y el tiempo de procesamiento.
- **Salidas:** *Prompts* optimizados junto con información sobre su calidad, desempeño lingüístico y nivel de adecuación contextual. Estos resultados sirven como insumo para la evaluación y validación de la hipótesis de investigación.

El modelo conceptual se alinea con las seis etapas metodológicas definidas en este trabajo:

1. Recolección y preparación de datos.
2. Generación y clasificación de *Prompts*.
3. Mejora y definición de métricas y parámetros de evaluación.
4. Implementación experimental del sistema.
5. Evaluación y validación de calidad y desempeño.
6. Confirmación de la hipótesis de investigación.

En conjunto, el modelo conceptual y el diagrama de flujo de datos (Figura 3.2) permiten visualizar de forma clara la interacción entre los componentes del sistema, identificando entradas, procesos, salidas y puntos de control. Además, garantizan la trazabilidad de los resultados, desde la formulación inicial de los *Prompts* hasta la validación final de la hipótesis, asegurando consistencia entre los aspectos cualitativos (claridad, coherencia, adecuación lingüística) y las métricas cuantitativas de desempeño (coherencia, relevancia y diversidad).

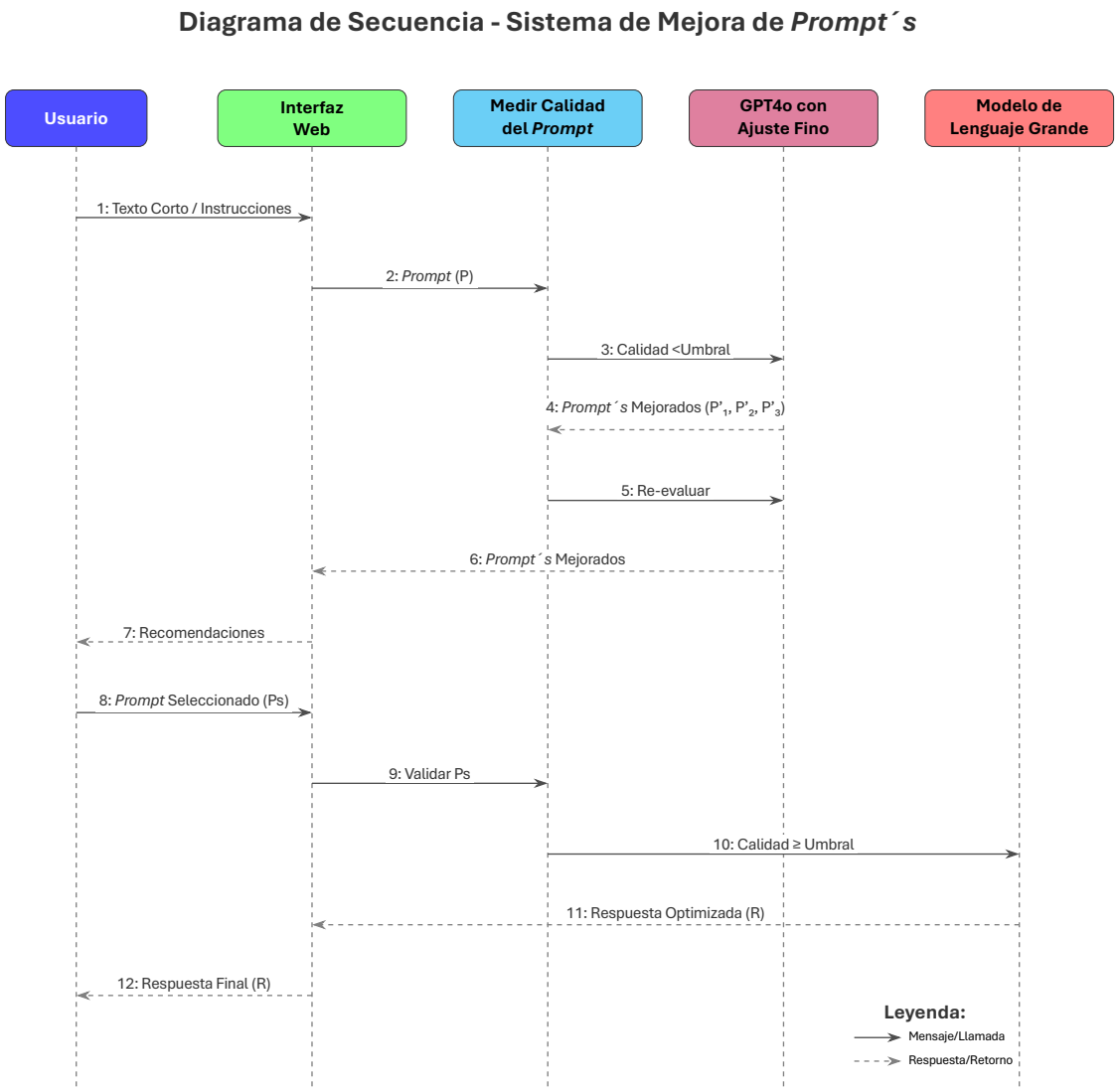

Figura 3.2: Diagrama de Flujo de Datos del sistema de mejora de *Prompts*.

3.4 Diseño de Componentes y Técnicas

Esta sección describe los componentes funcionales y las técnicas empleadas en el desarrollo del sistema de mejora de *Prompts*, alineados con las etapas metodológicas y el modelo conceptual previamente definidos. Cada módulo contribuye al flujo general de procesamiento mostrado en la Figura 3.2, asegurando la coherencia entre las entradas, el procesamiento y las salidas del sistema.

El diseño propuesto integra cinco componentes principales que interactúan de manera jerárquica para permitir la captura, evaluación, optimización y validación de los *Prompts*, garantizando tanto la calidad lingüística como el desempeño técnico del modelo de lenguaje empleado.

3.4.1 Módulo de Obtención y Anotación de *Prompts*

Este módulo constituye el punto de entrada del sistema y permite la introducción de *Prompts* sin procesar mediante una interfaz web. Los usuarios pueden ingresar instrucciones o consultas textuales, que son registradas y clasificadas según su tipo, complejidad y propósito comunicativo.

3.4.2 Indicador de Calidad (*Métrica de Evaluación*)

La evaluación de calidad de los *Prompts* se realiza mediante un indicador compuesto que integra métricas cuantitativas y cualitativas, permitiendo determinar si un *Prompt* cumple con el umbral de calidad requerido para ser aceptado o debe pasar por el proceso de mejora iterativa.

3.4.3 Técnica de Mejora de *Prompts*

La técnica de mejora se basa en la ejecución iterativa de un algoritmo de optimización lingüística sobre el modelo GPT-4o, configurado como sistema base. El proceso parte de un *Prompt* inicial (P) y genera versiones alternativas (P'_1, P'_2, P'_3), aplicando transformaciones de redacción, reordenamiento sintáctico y expansión contextual. Cada versión es revaluada por el módulo de calidad y el ciclo se repite hasta alcanzar el umbral mínimo establecido. Este procedimiento combina elementos de aprendizaje por refuerzo supervisado supervisado con reglas heurísticas de mejora lingüística, buscando equilibrio entre naturalidad, precisión y eficiencia del LLM.

3.4.4 Interfaz Web y Flujo de Consulta

La implementación del prototipo funcional se desarrolló como una aplicación web modular que integra los componentes anteriores en un flujo de consulta unificado. La interfaz permite

al usuario:

a) Ingresar *Prompts* manualmente

b) Recibir recomendaciones automáticas de mejora

c) Seleccionar ejemplos predefinidos

d) Consultar las respuestas generadas por el modelo tras la optimización

3.4.5 Preparación para la Experimentación

En esta fase se establecen las condiciones necesarias para la validación empírica del sistema. Se definen los conjuntos de prueba, las configuraciones del LLM, los criterios de control y las métricas comparativas que permitirán evaluar la efectividad del proceso de mejora. Asimismo, se documentan los parámetros de tiempo de respuesta, tamaño de *Prompts* y tamaño de las respuestas generadas por un LLM.

Estos datos servirán para analizar la correlación entre la mejora lingüística y las métricas de desempeño cuantitativo, consolidando la validez del modelo propuesto.

Capítulo 4

Experimentación y Resultados

Este capítulo presenta la ejecución práctica de la metodología propuesta para la mejora de *Prompts* descrita en el Capítulo 3, abarcando las etapas de implementación, evaluación y validación del sistema diseñado conforme al diagrama de flujo de datos mostrado en la Figura 3.2.

Los experimentos realizados se enfocan en el proceso completo de mejora lingüística de los *Prompts*, partiendo de la evaluación inicial mediante la métrica de calidad propuesta. Cada *Prompt* es analizado y comparado con el umbral establecido para determinar su nivel de calidad; en caso de no cumplirlo, se aplica el proceso de mejora que genera tres versiones optimizadas del *Prompt* original. Posteriormente, se evalúa el desempeño del LLM, considerando el tamaño del *Prompt*, la extensión de la respuesta generada y el tiempo de respuesta del modelo LLM.

La experimentación se llevó a cabo en un entorno controlado, integrando herramientas de NLP, análisis estadístico y visualización de datos. En esta fase se aplicaron criterios de control específicos como la longitud, claridad y estructura de los *Prompts*.

Los resultados experimentales demuestran la efectividad del sistema de mejora propuesto, al evidenciar aumentos sustanciales en la precisión lingüística de las respuestas del LLM y una reducción considerable en los tiempos de procesamiento. Estos hallazgos respaldan la solidez del enfoque metodológico y su aplicabilidad en escenarios reales de generación de *Prompts*.

4.1 Entorno de Desarrollo

El desarrollo de los experimentos se llevó a cabo en un entorno mixto que combinó herramientas de análisis estadístico, NLP y ajuste fino a un LLM. Se emplearon los lenguajes Python y R, utilizando entornos de ejecución en la nube mediante Google Colab para pruebas rápidas, aunque la mayor parte del trabajo se realizó en un entorno local para asegurar control total

sobre los recursos y los datos.

Los experimentos se enfocaron en evaluar y mejorar la calidad lingüística de los *Prompts*, midiendo variables como la extensión del *Prompt*, la longitud de la respuesta generada, el tiempo de respuesta del LLM y el cumplimiento de un umbral de calidad determinado mediante la métrica propuesta. Esto permitió analizar cuantitativa y cualitativamente el desempeño del sistema antes y después de la aplicación de las mejoras.

4.1.1 Hardware y Software

El entorno local utilizado para los experimentos estuvo configurado con las siguientes características de hardware:

- **Procesador:** Intel Core i5 a 2.50 GHz
- **Memoria RAM:** 16 GB
- **Memoria GPU:** 4 GB
- **Sistema operativo:** Windows 11

Las principales librerías y *frameworks* empleados fueron:

- **Transformers** y **OpenAI API**: interacción con GPT-4o y generación/mejora de *Prompts*.
- **NLTK** y **spaCy**: NLP y análisis sintáctico.
- **Pandas** y **NumPy**: manejo y análisis de datos.
- **Scikit-learn**: análisis multivariable y técnicas estadísticas.
- **Matplotlib** y **Seaborn**: visualización de resultados.
- **Torch**: entrenamiento de LLMs.
- **Datasets**: manejo de corpus.

Esta configuración permitió automatizar la evaluación de los *Prompts*, generar versiones mejoradas, medir la calidad lingüística mediante la métrica propia y analizar el desempeño del LLM considerando tiempo de respuesta, tamaño del *Prompt* y extensión de la respuesta generada.

4.1.2 Crear Corpus en Español Culto

El principal propósito de este corpus es facilitar el proceso de ajuste fino del LLM para que se especialice en el dominio de la historia de la Ciudad de México. Al entrenar el LLM con este corpus, se busca optimizar su capacidad para generar respuestas optimizadas sobre temas históricos relacionados con la Ciudad de México, aplicando un lenguaje culto y adecuado para audiencias académicas o interesadas en el estudio profundo de la historia de la ciudad.

Este corpus no solo ayudará a mejorar la calidad de las respuestas del LLM, sino que también permitirá evaluar su capacidad para comprender y manejar el contexto cultural y lingüístico mexicano en diversas situaciones históricas, lo que resulta esencial para la aplicación del LLM en investigaciones académicas y educativas.

Basado en contenido textual de la revista “Arqueología Mexicana”, en su edición especial 115 titulada “CDMX Guía de Viajeros (Parte 1)” (Arqueología Mexicana, 2024). Esta revista cuya portada se ilustra en la Figura 4.1, se caracteriza por ofrecer información detallada y rigurosa sobre la historia y cultura de la Ciudad de México, sirvió como fuente primaria para construir un conjunto de datos que abarca una variedad de temas relevantes, adaptados al contexto histórico y cultural de la capital mexicana.

Figura 4.1: Portada de la revista *Arqueología Mexicana*.

Diversidad de Temas Lingüísticos

El corpus no sólo se centró en la variedad de contenido histórico, sino también en una diversidad de temas lingüísticos que cubren aspectos clave del lenguaje. Estos temas permiten mejorar la capacidad del modelo para manejar una gama amplia de estructuras y dificultades lingüísticas. Los 16 temas lingüísticos aplicados en el corpus fueron:

- Frases Afirmativas
- Declaraciones Incompletas o Frases Sin Estructura de Pregunta
- Frases Tipo Pregunta (pero mal formuladas)
- Ambigüedad
- Contexto
- Formalidad y Tono
- Redundancias
- Concordancia (sujeto-verbo)
- Puntuación (falta o exceso)
- Uso Incorrecto de Pronombres
- Uso Incorrecto de Preposiciones
- Excesiva Especificidad o Falta de Generalización
- Errores de Traducción
- Uso Incorrecto de Tiempos Verbales
- Errores de Capitalización
- Vocabulario

Cada uno de estos temas fue diseñado para abordar diferentes retos del lenguaje y asegurar que el LLM pudiera generar respuestas optimizadas a los distintos estilos y registros del español culto mexicano.

Arquitectura del corpus
En búsqueda de crear un corpus que cumpla con los criterios de arquitectura empleados en lingüística, a continuación, se definen los elementos esenciales considerados durante su elaboración.

Autenticidad: Las oraciones y fragmentos textuales extraídos directamente de la revista *Arqueología Mexicana* mantienen la autenticidad del discurso académico y divulgativo. Este material refleja un uso genuino del español culto, empleado por especialistas en historia, arqueología y cultura mexicana.

Representatividad: El corpus ofrece una representación significativa de la variedad temática y lingüística presente en textos históricos y culturales de carácter académico. Abarca contenidos relacionados con la evolución urbana, el arte, la religión y la vida cotidiana en la Ciudad de México, lo que asegura una cobertura equilibrada de distintos contextos históricos.

Diversidad o variedad: Las oraciones seleccionadas presentan una diversidad notable en términos de vocabulario especializado, estructuras gramaticales complejas y estilos narrativos. Esto permite que el modelo aprenda distintos niveles de formalidad, profundidad conceptual y riqueza léxica característicos del español culto mexicano.

Tamaño: El corpus compuesto por 2000 oraciones constituye una base sustancial para el entrenamiento del LLM. Si bien no es exhaustivo, su tamaño es suficiente para realizar análisis lingüísticos representativos y obtener patrones relevantes del uso del español académico

e histórico.

Anotación: Para garantizar un análisis preciso, las oraciones fueron clasificadas según su función dentro del corpus: *preguntas, respuestas y contextos explicativos*. Esta organización facilita el proceso de ajuste fino del LLM, permitiendo una mejor comprensión del vínculo entre la formulación de *Prompts* y las respuestas generadas.

Estructura del Corpus

El corpus creado a partir de la revista, fue estructurado de manera que facilitara el ajuste fino del LLM. Para ello, se organizó la información como se muestra en la Tabla 4.1 con cuatro columnas clave que se describen a continuación:

- **Categoría**: La categoría se refiere al tipo de enfoque o profundidad del contenido presentado en el *Prompt*.
- **Pregunta**: La pregunta corresponde a una consulta formulada en relación con un tema histórico de la Ciudad de México.
- **Respuesta**: La respuesta ofrece una explicación detallada sobre el tema, en base al contenido de la revista.
- **Contexto**: El contexto contextualiza la respuesta, proporcionando detalles adicionales que permiten comprender mejor la importancia del tema en la historia de la Ciudad de México.

Tabla 4.1. Ejemplos de estructura del corpus formulado a partir de *Prompts* históricos.

Categoría	Pregunta	Respuesta	Contexto
Variedad de Preguntas	¿Cuáles fueron las principales características del arte renacentista?	El arte renacentista se caracteriza por el uso de la perspectiva, el realismo en la representación de la figura humana, la búsqueda de la belleza ideal y la influencia de la cultura clásica grecorromana.	El Renacimiento, que se desarrolló entre los siglos XIV y XVII en Europa, marcó un resurgimiento del interés por la antigüedad clásica, lo que se reflejó en la pintura, la escultura y la arquitectura de la época.
Profundidad del Contexto	¿Qué papel desempeñó la Calle José María Pino Suárez en la planificación urbana de la Ciudad de México?	La Calle José María Pino Suárez desempeñó un papel crucial en la planificación urbana, siendo uno de los ejes que estructuraron tanto a Tenochtitlan prehispánica como a la ciudad virreinal, facilitando el comercio y la comunicación.	Ha sido fundamental para conectar diferentes áreas de la ciudad, lo que ha permitido su desarrollo a lo largo de los siglos, adaptándose a las necesidades cambiantes de la población.

Tabla 4.1. Ejemplos de estructura del corpus formulado a partir de *Prompts* históricos (continuación).

Categoría	Pregunta	Respuesta	Contexto
Complejidad de las Respuestas	¿Qué implicaciones tiene la representación artística de San Lucas para la percepción de la fe en la sociedad contemporánea?	La representación artística de San Lucas puede influir en la percepción de la fe al ofrecer un modelo de compasión y dedicación, pero también plantea interrogantes sobre la relevancia de las figuras religiosas en un mundo cada vez más secular.	En la actualidad, las representaciones de figuras religiosas como San Lucas pueden ser vistas como un puente entre la tradición y la modernidad, invitando a la reflexión sobre el papel de la espiritualidad en la vida cotidiana de las personas.

4.1.3 Selección del LLM

En esta sección se presentan los LLMs considerados para el proyecto, seleccionados a partir de la revisión bibliográfica (Tabla 4.2) y de la exploración de LLMs disponibles en *Hugging Face* (Hugging Face, s.f.), como se muestra en la Tabla 4.3.

Nuestra investigación mostró que modelos como BERT y RoBERTa, aunque sobresalen en tareas de análisis, comprensión semántica y clasificación, requieren plantillas predefinidas y se orientan principalmente a la traducción de texto, lo cual limita su aplicabilidad para la generación libre de texto que demanda este proyecto.

Por ello, se priorizaron modelos capaces de generar texto de manera autónoma, seleccionando GPT-2, GPT-4, LLaMA y Mistral, clasificados como generativos en *Hugging Face*, garantizando así capacidad generativa, comprensión del contexto en español y factibilidad técnica.

Tabla 4.2. Modelos revisados en la literatura

Modelo	Referencia	Tipo	Tarea principal
BERT	(Han et al., 2021); (Amur et al., 2023)	Enmascarado	Similitud semántica, comprensión de contexto y pregunta y respuesta
RoBERTa-large	(Min et al., 2023)	Enmascarado	Extracción semántica y tareas proxy de pregunta y respuesta
GPT-2	(McCoy et al., 2023); (Singh et al., 2023)	Generativo	Generación de texto y modelado semántico complejo
GPT-3	(Singh et al., 2023)	Generativo	Generación de texto y completado de *Prompts*

Tabla 4.3. Modelos explorados en *Hugging Face*

Modelo	Referencia	Tipo	Descripción
GPT-2	(Hugging Face, s.f.)	Generativo	Probado para generación de *Prompts*; buen desempeño en español limitado
LLaMA	(Hugging Face, s.f.)	Generativo	Mayor requerimiento de hardware y menor manejo del contexto en español
Mistral	(Hugging Face, s.f.)	Generativo	Rendimiento prometedor, pero las pruebas iniciales mostraron limitaciones técnicas
GPT-4o	(OpenAI, s.f.)	Generativo	Seleccionado por su alto rendimiento, comprensión contextual y disponibilidad vía API

Justificación de la selección final del LLM

Si bien BERT y RoBERTa se destacan en la literatura por su eficacia en tareas de análisis semántico y detección de similitud textual, no corresponden a LLMs de tipo generativo.

Dado que el objetivo del experimento es evaluar técnicas de mejora de *Prompts* en la generación de texto, se priorizaron LLMs explícitamente generativos: GPT-2, LLaMA y Mistral, clasificados como tales en el *Hugging Face Hub* (Hugging Face, s.f.).

En el caso de GPT-4o, este modelo no se obtuvo del repositorio de *Hugging Face*, sino que se considera la evolución directa de GPT-3, ampliamente documentada en la literatura y accesible mediante la API oficial de OpenAI (OpenAI, s.f.). Su inclusión se justifica por su rendimiento superior en comprensión contextual, coherencia generativa y manejo del idioma español.

Entre los LLMs evaluados, GPT-4o fue finalmente seleccionado como el LLM base para la versión final del sistema, debido a su escalabilidad, estabilidad en entornos de producción y disponibilidad mediante infraestructura en la nube. Además, este modelo se sometió a un proceso de entrenamiento adicional orientado al dominio temático de la historia de la Ciudad de México, con el fin de optimizar la coherencia, precisión léxica y adecuación estilística en las respuestas generadas.

Este enfoque permite aprovechar recursos computacionales distribuidos y optimizados por OpenAI, eliminando la necesidad de hardware local y garantizando tiempos de respuesta consistentes en los procesos de generación de texto, de acuerdo con los objetivos de desempeño y sostenibilidad del proyecto.

4.2 Optimización y Evaluación de *Prompts*

La optimización y evaluación de *Prompts* constituye una fase esencial en el diseño de sistemas basados en LLMs. Su propósito es mejorar la formulación de las entradas para obtener respuestas más precisas y coherentes.

Este proceso permite establecer criterios de calidad y definir parámetros que regulan el comportamiento del LLM ante diferentes tipos de instrucciones, favoreciendo un control más eficiente sobre los resultados generados.

4.2.1 Técnica de Mejora de *Prompts*

La mejora iterativa de *Prompts* es un proceso clave para optimizar la interacción con LLMs. Este enfoque permite refinar *Prompts* iniciales mediante iteraciones guiadas, asegurando claridad y corrección gramatical. Utilizando un LLM, el algoritmo integra contexto y parámetros personalizados, generando respuestas más adecuadas al público objetivo. Este método es esencial para aplicaciones académicas o creativas, por ejemplo. En el Algoritmo 2 se muestra la propuesta para la mejora de *Prompts*.

El algoritmo de mejora iterativa de *Prompts* toma un texto inicial y lo refina a través de un LLM. Utiliza una combinación de contexto, instrucciones específicas y ajustes en los parámetros del modelo para generar versiones mejoradas del *Prompt* en varias iteraciones. El resultado final es un texto mejorado en claridad, corrección y relevancia para el público objetivo, logrando así una mayor efectividad en su aplicación.

Algoritmo 2: Proceso Integral de Mejora y Evaluación de *Prompts*.

Input: $prompt\_usuario, contexto, umbral\_calidad, max\_iteraciones$
Output: $prompt\_optimizado, respuesta\_final$
▷ `Fase 1: Recepción del Prompt`
Recibir $prompt\_usuario$ desde la Interfaz Web
Mostrar instrucciones y contexto al usuario

▷ `Fase 2: Evaluación inicial de calidad`
$calidad \leftarrow medir\_calidad(prompt\_usuario)$
if $calidad \geq umbral\_calidad$ **then**
    Enviar $prompt\_usuario$ al Modelo de Lenguaje (LLM)
    $respuesta\_final \leftarrow LLM(prompt\_usuario)$
    return $prompt\_usuario, respuesta\_final$
end

▷ `Si no cumple con el umbral, continuar en la siguiente parte.`

Algoritmo 2: Proceso Integral de Mejora y Evaluación de *Prompts* (continuación).

```
▷ Fase 3: Proceso iterativo de mejora
prompt_actual ← prompt_usuario
for i ← 1 to max_iteraciones do
    prompt_nuevo ← GPT4o_ajustado(prompt_actual, contexto)
    calidad ← medir_calidad(prompt_nuevo)
    if calidad ≥ umbral_calidad then
        prompt_optimizado ← prompt_nuevo
        break
    end
    prompt_actual ← prompt_nuevo
end
▷ Fase 4: Validación y entrega de resultados
if calidad < umbral_calidad then
    Mostrar recomendaciones al usuario mediante la Interfaz Web
    prompt_optimizado ← prompt_actual
end
respuesta_final ← LLM(prompt_optimizado)
Enviar respuesta_final y prompt_optimizado al usuario
return prompt_optimizado, respuesta_final
```

La estrategia propuesta para mejorar los *Prompts* fue implementada utilizando el LLM GPT-4o de OpenAI, accesible a través de su API. Este LLM, parte de la familia GPT-4, representa una arquitectura optimizada multimodal, capaz de procesar texto, imágenes y audio de manera unificada. Aunque es un modelo generalista, GPT-4o ha demostrado una notable capacidad de comprensión y generación de lenguaje en español, lo cual lo hace altamente adecuado para la mejora iterativa de *Prompts* en dicho idioma.

Características del modelo GPT-4o

- **Arquitectura y capacidades:** GPT-4o ("o" de "omni") es un modelo unificado que procesa entrada y salida en múltiples modalidades. A pesar de esta complejidad, su desempeño en tareas puramente textuales supera ampliamente a generaciones anteriores como GPT-3.5, y su tiempo de respuesta ha sido optimizado, siendo más rápido y eficiente.

- **Entrenamiento:** GPT-4o fue entrenado con un corpus masivo y multilingüe que incluye una amplia representación del idioma español. Esto le otorga la capacidad de entender y generar texto con alta calidad en diferentes variantes del español, incluyendo el español culto utilizado en contextos académicos o históricos.

- **Tokenización:** Utiliza una tokenización unificada y optimizada, basada en el enfoque Codificación por Pares de Bytes a nivel de *byte* (del inglés *Byte-Level Byte Pair Encoding*, BPE), con un vocabulario capaz de capturar patrones lingüísticos en múltiples idiomas. Esta tokenización permite una representación precisa del contenido semántico y morfosintáctico, incluso en lenguas con estructuras complejas como el español.

Proceso de Implementación
Se partió del modelo base GPT-4o y se procedió a su adaptación mediante un proceso de ajuste fino, utilizando ejemplos de textos en español culto centrados en el contexto histórico de la Ciudad de México.

Esta etapa de ajuste permitió especializar las respuestas del LLM para ofrecer resultados más precisos y coherentes en dicho dominio. Además, se complementó el entrenamiento con técnicas de Ingeniería de *Prompts*, aprovechando las capacidades del LLM para interpretar instrucciones complejas a través de estrategias *few-shot* y *zero-shot prompting*.

4.2.2 Métrica de Calidad de *Prompts*

Para evaluar la calidad de los *Prompts* en la interacción con los LLMs, se propone la métrica PromptIQ. Esta métrica permite analizar y optimizar aspectos clave de los *Prompts*, como claridad, coherencia y precisión. Su aplicación fomenta la generación de respuestas más efectivas, aumentando la aplicabilidad de los LLMs en diversos contextos y mejorando su rendimiento en tareas específicas.

La métrica PromptIQ se modela mediante regresión lineal multivariable, donde cada característica del *Prompt* (claridad, coherencia y precisión) se asocia a un coeficiente que indica su impacto en la calidad general. En esta ecuación, el valor constante 0.9987 corresponde al *intercepto*, que representa el valor base de PromptIQ cuando todos los errores o características medibles son cero. Los coeficientes de cada variable independiente muestran cómo cada error lingüístico disminuye o afecta el valor final de la métrica.

Con esta formulación, el Algoritmo 3 permite calcular los coeficientes de la métrica PromptIQ y generar versiones mejoradas de los *Prompts* cuando el valor estimado cae por debajo del umbral deseado.

Algoritmo 3: Cálculo de los coeficientes de la métrica PromptIQ mediante regresión lineal.

Input: Prompt P inicial
Output: Prompt válido o tres Prompts mejorados
Function `CalcularPromptIQ(`P`)`**:**
  Enviar P a GPT-4 para contabilizar los errores lingüísticos ; `// Obtener conteos individuales`
  Recibir $\alpha, \omega, \zeta, \gamma, \phi, \lambda, \rho, \tau, \chi, \kappa$; `// Errores por coeficiente`
  Calcular Y**:**

$$Y = 0.9987 - 0.018 \cdot \alpha - 0.0145 \cdot \omega - 0.0904 \cdot \zeta - 0.0458 \cdot \gamma - 0.0908 \cdot \phi - 0.0453 \cdot \lambda - 0.0454 \cdot \rho - 0.0919 \cdot \tau - 0.0902 \cdot \chi - 0.0885 \cdot \kappa - 0.0914 \cdot \varepsilon$$

Algoritmo 3: Cálculo de los coeficientes de la métrica PromptIQ mediante regresión lineal (continuación).

```
if Y ≥ umbral_α then
    return P ;                                          // Prompt válido
else
    Mejorar P usando un modelo finetuneado en el dominio específico
    Generar 3 Prompts mejorados: P_1, P_2, P_3
    return P_1, P_2, P_3 ;                              // Prompts mejorados
```

Este algoritmo está enfocado en el análisis estadístico de regresión lineal multivariable, como se describe en la Sección 2.5.1, el cual permite modelar la relación entre diversas características del *Prompt*. La regresión lineal multivariable es una técnica estadística utilizada para modelar la relación entre una variable dependiente (o variable de respuesta) y múltiples variables independientes (o predictoras).

En el contexto de la evaluación de la calidad de los *Prompts*, este análisis se emplea para entender cómo diferentes características de un *Prompt* (como claridad, coherencia, precisión, etc.) impactan en las métricas de calidad del texto generado, como la gramaticalidad, relevancia, entre otras.

En la Sección 2.5.1 se pueden consultar a detalle los conceptos clave del análisis de regresión lineal multivariable.

Ecuación del modelo.
La ecuación del modelo sigue la formulación descrita en la Ecuación 2.13 correspondiente a la Sección 2.5.1.

Interpretación de los coeficientes.
Los coeficientes (β) indican la fuerza y la dirección de la relación entre las variables independientes y la variable dependiente. Si el coeficiente de una variable independiente es positivo, significa que un aumento en esa variable independiente está asociado con un aumento en la variable dependiente. Si es negativo, significa que un aumento en la variable independiente está asociado con una disminución en la variable dependiente.

La formulación matemática para calcular la métrica PromptIQ se muestra en la Ecuación 4.1:

$$\begin{aligned}\hat{Y} = \; & 0.9987 - 0.018 \cdot \alpha - 0.0145 \cdot \omega - 0.0904 \cdot \zeta \\ & - 0.0458 \cdot \gamma - 0.0908 \cdot \phi - 0.0453 \cdot \lambda \\ & - 0.0454 \cdot \rho - 0.0919 \cdot \tau - 0.0902 \cdot \chi \\ & - 0.0885 \cdot \kappa - 0.0914 \cdot \varepsilon \end{aligned} \tag{4.1}$$

(con $\varepsilon = 0$, cuando no hay error).

Por otro lado, cada coeficiente independiente de la Ecuación 4.1 está representado por una letra griega que se puede visualizar en la Tabla 4.4.

Tabla 4.4. Correspondencia entre errores lingüísticos y letras griegas

Término	Letra griega	Nombre de la letra griega
Puntuación	α	Alfa
Ortografía	ω	Omega
Sintaxis	ζ	Zeta
Gramática	γ	Gamma
Fluidez	ϕ	Fi
Léxico	λ	Lambda
Registro	ρ	Rho
Pragmatismo	τ	Tau
Contexto	χ	Ji (Chi)
Incoherencia	κ	Kappa
Semántica	ε	Épsilon

4.3 Definición de Parámetros de *Prompts*

En esta sección se presentan los parámetros fundamentales que permiten evaluar y optimizar los *Prompts* utilizados con los LLMs. La correcta definición de estos parámetros asegura una generación de respuestas más precisa, coherente y relevante, además de facilitar la medición cuantitativa de la calidad de los *Prompts*.

4.3.1 Umbral de Calidad de un *Prompt*

Este análisis estadístico se llevó a cabo a partir de un conjunto total de 3,592 *Prompts*, conformado por 1,796 *Prompts* originales y sus correspondientes 1,796 versiones tipo pregunta. Los datos fueron extraídos del foro `/preguntaleareddit`, disponible públicamente en la plataforma https://www.reddit.com.

Estos Prompts corresponden a la segunda extracción descrita en la Sección 3.2.1, que consolidó el conjunto completo de datos para análisis estadísticos. Los textos fueron transformados automáticamente en documentos tipo pregunta mediante el LLM GPT-4o, con el objetivo de normalizar su estructura y facilitar su evaluación posterior bajo criterios de calidad.

Con el objetivo de establecer un criterio cuantitativo que permita determinar si un *Prompt* posee calidad suficiente o no, se llevó a cabo un experimento comparativo entre tres técnicas comúnmente empleadas en análisis de distribución de datos: el uso del centroide del grupo más numeroso (según agrupación $k-means$), la Estimación de Densidad de *Kernel* , y el Análisis de Frecuencias por Histograma.

Justificación breve de cada método:

- $\boldsymbol{k-means}$**:** identifica el centroide del grupo dominante de valores PromptIQ, representando la tendencia central de *Prompts* con calidad aceptable.
- **KDE:** estima la densidad continua de valores PromptIQ, mostrando dónde se concentra la mayoría de *Prompts* con alta calidad.
- **Histograma**: proporciona un enfoque simple para detectar la frecuencia más alta de valores PromptIQ de los *Prompts*, útil como referencia rápida.

Cada una de estas técnicas permitió obtener un valor umbral representativo del comportamiento global de la métrica PromptIQ, la cual funge como indicador central de desempeño.

Como se observa en la Tabla 4.5, los valores obtenidos para el umbral de calidad varían ligeramente entre técnicas: el centroide del grupo más grande según $k-means$ arrojó un umbral de 0.846, la técnica de KDE identificó el valor de máxima densidad en 0.844, mientras que el histograma reveló una mayor frecuencia en el valor 0.850. Estas ligeras diferencias reflejan la sensibilidad de cada enfoque ante la distribución subyacente de los datos.

Tabla 4.5. Valores de diferentes umbrales de calidad PromptIQ aplicando técnicas estadísticas.

Técnica Estadística	Valor del Umbral de Calidad (PromptIQ)
$k-means$	0.846
KDE	0.844
Histograma	0.850

Complementariamente, la técnica de agrupamiento $k-means$ permitió segmentar los datos en función de su similitud, considerando tanto el valor PromptIQ como el tiempo de respuesta, como se muestra en la Figura 4.2. El centroide del grupo con mayor población se utilizó como valor representativo del umbral, asumiendo que esta agrupación refleja la clase dominante de *Prompts* de calidad aceptable. Sin embargo, dicha técnica está condicionada por la inicialización de los centroides y puede ser sensible a la forma no esférica de los datos.

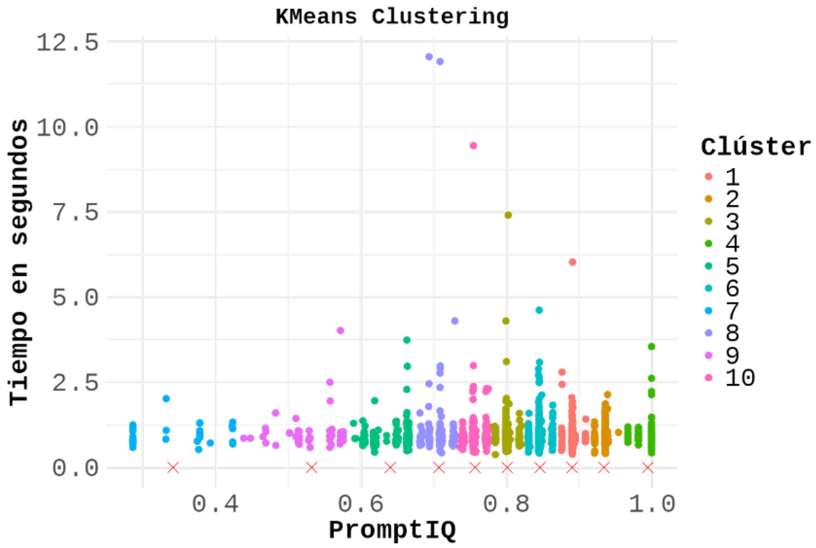

Figura 4.2: Segmentación $k-means$ de los *Prompts* basada en PromptIQ y Tiempo de Respuesta.

Cabe destacar que, entre las tres técnicas evaluadas, KDE resulta particularmente ventajosa por su capacidad de generar una visualización continua y suavizada de la distribución de los datos, sin depender de particiones arbitrarias como ocurre en los histogramas.

Esta técnica permite observar la estructura real de los datos sin los sesgos inducidos por la elección de ancho de banda de los intervalos, lo cual se evidencia claramente en la Figura 4.3, donde se identifica con precisión el valor de PromptIQ con mayor densidad relativa. Este valor representa, por tanto, el punto de mayor concentración de calidad de *Prompts* según la métrica evaluada.

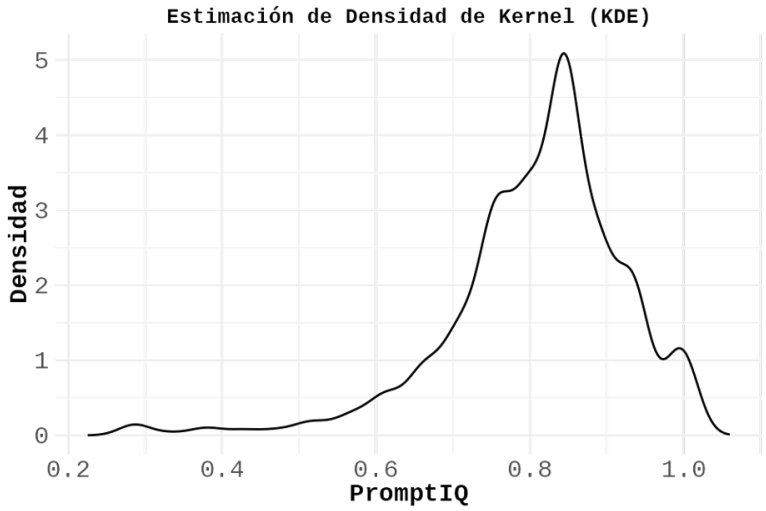

Figura 4.3: Distribución de PromptIQ mediante Estimación de Densidad de Kernel.

Por otro lado, el análisis de frecuencias mediante histograma, ilustrado en la Figura 4.4, proporcionó un enfoque más directo para estimar el umbral mediante la observación de la barra con mayor altura. Aunque es sencillo de implementar, este método es susceptible al número y tamaño de los intervalos definidos, lo cual puede afectar la precisión de la estimación.

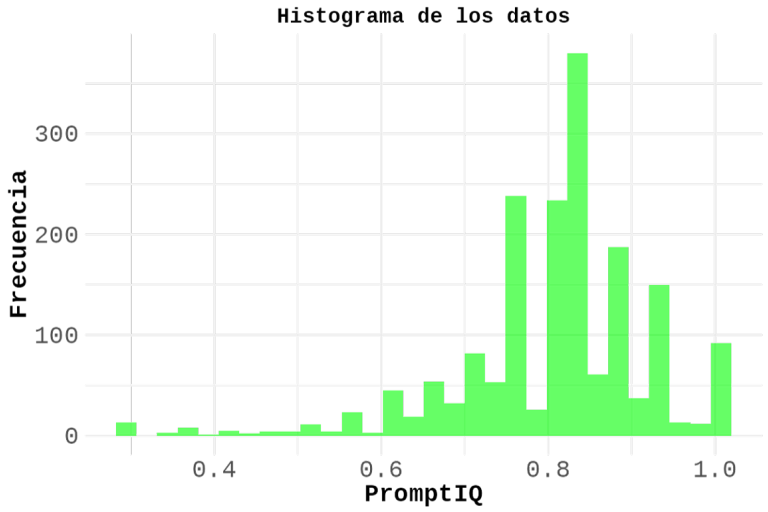

Figura 4.4: Distribución de Frecuencia de PromptIQ.

En conjunto, estos tres enfoques permitieron contrastar distintos criterios para definir el umbral de calidad, resaltando la robustez y adaptabilidad de KDE como técnica descriptiva. Por su naturaleza no paramétrica y su capacidad para modelar distribuciones complejas, KDE se posiciona como la opción más confiable en contextos donde se busca identificar patrones reales en los datos sin supuestos rígidos.

En consecuencia, se establece un umbral de calidad 0.844; los *Prompts* con valores iguales o superiores a este pueden considerarse de alta calidad, mientras que aquellos con valores inferiores se clasifican como de calidad baja o insuficiente.

4.3.2 Tamaño de un *Prompt*

Este análisis estadístico se llevó a cabo a partir de un conjunto total de 3,592 *Prompts*, conformado por 1,796 versiones originales extraídas del foro `/preguntaleareddit`, disponible públicamente en la plataforma https://www.reddit.com, y 1,796 versiones tipo pregunta generadas posteriormente mediante el LLM GPT-4o.

Estos *Prompts* corresponden a la segunda extracción descrita en la Sección 3.2.1, que consolidó el conjunto completo de datos para los análisis estadísticos. Las versiones tipo pregunta fueron creadas a partir de los textos originales con el objetivo de normalizar su estructura, mejorar su claridad lingüística y facilitar su evaluación posterior bajo criterios de calidad.

Para orientar el análisis, se verificó si la variable "número de palabras" seguía una distribución normal mediante la prueba de Shapiro-Wilk cuyos resultados se muestran en la Tabla 4.6. Dado que los datos no cumplieron con la normalidad, se emplearon técnicas descriptivas robustas: $k-means$, Estimación de Densidad de *Kernel* y Análisis de Frecuencias por Histograma.

Estas técnicas fueron seleccionadas por su capacidad de:

- $\boldsymbol{k-means}$**:** permite identificar el centroide del grupo dominante de *Prompts*, proporcionando un valor de referencia para el "número de palabras" óptimo.
- **KDE:** permite estimar la densidad continua de los *Prompts*, mostrando dónde se concentra la mayoría y capturando la estructura real de los datos.
- **Histograma**: permite detectar las frecuencias absolutas de "número de palabras", ofreciendo un enfoque simple y directo para referencia rápida.

Esta estrategia permitió describir adecuadamente los datos y determinar un umbral representativo para el número mínimo de palabras.

Tabla 4.6. Prueba de normalidad en las variables "número de palabras" y PromptIQ.

Prueba de normalidad de Shapiro-Wilk	**Resultado**
W	0.6352
Valor p	$5.034447e-52$
Interpretación	Los datos NO siguen una distribución normal porque el valor p es menor que α (0.05).

Leyendas:

- W: Estadístico de Shapiro-Wilk.
- p: Valor p asociado.

Los resultados comparativos de los umbrales de número mínimo de palabras y su valor PromptIQ se presentan en la Tabla 4.7.

Tabla 4.7. Valores de diferentes umbrales de número mínimo de palabras y PromptIQ.

Método Estadístico	**Número de Palabras**	**PromptIQ**
$k-means$	11	0.8106
KDE	10	0.9349
Histograma	8	0.7109

El método estadístico de $k-means$, ilustrado en la Figura 4.5, permitió identificar grupos naturales de *Prompts* en función del "número de palabras" y del valor PromptIQ. El centroide del grupo con mayor población sirvió como referencia inicial, ubicando el umbral en 11 palabras con un PromptIQ de 0.8106.

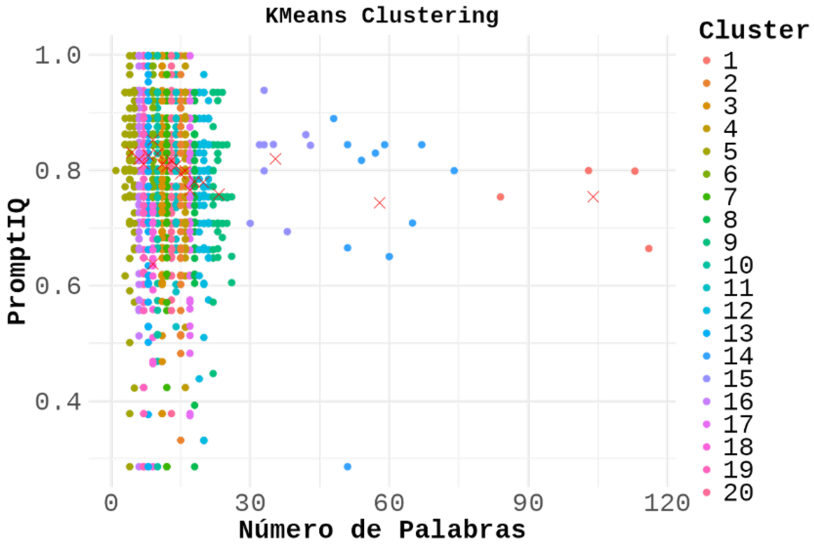

Figura 4.5: $k-means$ **Clustering de *Prompts* basado en PromptIQ y "número de palabras".**

La Estimación KDE, mostrada en la Figura 4.6, permitió capturar la forma real de la distribución de palabras sin asumir normalidad. El máximo de densidad indica un umbral de 10 palabras, asociado con un PromptIQ de 0.9349, el valor más alto entre las técnicas evaluadas.

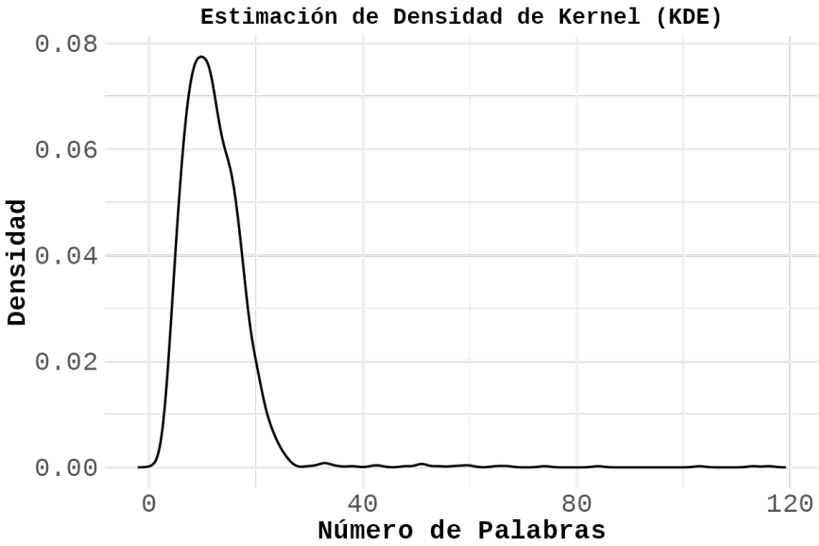

Figura 4.6: Distribución de "número de palabras" mediante Estimación de Densidad de Kernel.

Por último, el histograma Figura 4.7 mostró las frecuencias absolutas de cada "número de palabras", ubicando el umbral mínimo en 8 palabras con un PromptIQ de 0.7109. Este método proporciona una referencia rápida aunque más sensible a la elección de intervalos.

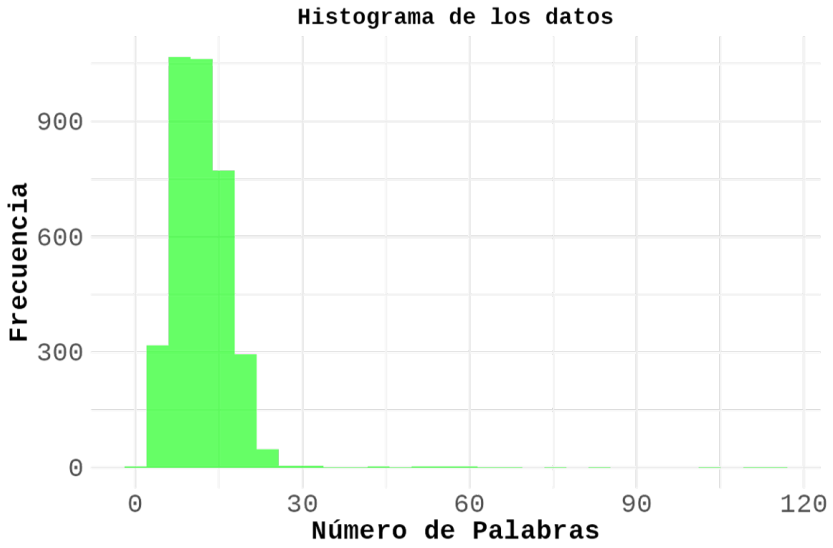

Figura 4.7: Distribución de Frecuencia de "número de palabras".

En conjunto, estos tres métodos estadísticos permitieron contrastar distintos criterios para determinar el número mínimo de palabras óptimo. El método KDE se destaca por su naturaleza no paramétrica, capaz de modelar distribuciones reales y señalar un umbral robusto de 10 palabras. Por ello, se adopta como referencia para mejorar la calidad de los *Prompts*.

4.3.3 Calcular el Número de Palabras Óptimo para Obtener Respuestas Claras por Parte de un LLM

En continuidad con la Etapa 1: Recolección y preparación de datos, esta sección aplica la metodología sobre un *dataset* especializado: el *Multilingual BioASQ-6B* en español. Este conjunto está conformado por un total de 2,251 *Prompts* tipo pregunta con sus respectivas respuestas, redactadas por expertos humanos en el ámbito biomédico.

La selección de este *dataset* responde a su alta calidad lingüística y conceptual, ya que garantiza coherencia semántica, precisión terminológica y una correspondencia clara entre las preguntas y las respuestas.

A diferencia de corpus abiertos como los de Reddit, donde el lenguaje y la veracidad pueden ser variables, el uso de este *dataset* permite evaluar de manera controlada y confiable la calidad esperada de las respuestas que un LLM debería generar, tomando como referencia las

redactadas por expertos humanos.

Esta elección facilita, además, el estudio del efecto del "número de palabras" como variable explicativa sobre las métricas BLEU, ROUGE_L, F_1-*Score* y Perplejidad.

Para identificar el rango de palabras que maximiza la calidad de las respuestas, se agruparon los datos en intervalos de "número de palabras" y se calcularon los valores promedio de cada métrica. La Tabla 4.8 presenta los resultados, resaltando los valores más altos y el número de registros por grupo, lo que facilita determinar rangos prácticos para la construcción de respuestas óptimas en un LLM.

Tabla 4.8. Valores de las métricas BLEU, ROUGE_L, $F_1 - Score$ y Perplejidad de acuerdo al rango de "número de palabras".

Rango Palabras	BLEU	ROUGE_L	$F_1 - Score$	Perplejidad	Registros
1–5	0.0178	0.0923	0.0797	1441.0360	53
6–10	0.2306	0.4524	0.4039	141.0885	123
11–15	**0.2153**	**0.4171**	**0.4162**	**87.2707**	**236**
16–20	**0.1744**	**0.3408**	**0.3661**	**50.9911**	**244**
21–25	**0.1336**	**0.2982**	**0.3219**	**46.2988**	**231**
26–30	0.1108	0.2634	0.2991	40.0912	189
31–35	0.0937	0.2407	0.2842	36.2955	157
36–40	0.0675	0.2079	0.2557	36.2949	110
41–45	0.0580	0.1733	0.2206	41.4068	102
46–50	0.0634	0.1808	0.2515	33.7509	96
51–55	0.0510	0.1698	0.2168	32.9636	77
56–60	0.0719	0.1790	0.2397	31.1247	57
61–65	0.0582	0.1709	0.2366	32.3626	57
66–70	0.0596	0.1614	0.2189	26.3600	58
71–75	0.0395	0.1447	0.2139	32.0350	47
76–80	0.0478	0.1379	0.2011	28.3005	32
81–85	0.0362	0.1363	0.1907	29.4245	52
86–90	0.0348	0.1160	0.1917	32.6163	27
91–95	0.0295	0.1195	0.1783	29.8024	30
96–100	0.0304	0.1107	0.1699	31.1842	30
101–105	0.0330	0.1208	0.1650	31.3034	21
106–110	0.0331	0.1197	0.1754	28.6241	26
111–115	0.0287	0.0966	0.1638	30.9796	35
116–120	0.0251	0.0894	0.1554	31.4584	14
121–125	0.0272	0.1048	0.1601	32.7222	19
126–130	0.0237	0.1095	0.1695	25.2483	15

Tabla 4.8. Valores de las métricas BLEU, ROUGE_L, $F_1 - Score$ y Perplejidad de acuerdo al rango de "número de palabras" (*continuación*).

Rango Palabras	BLEU	ROUGE_L	$F_1 - Score$	Perplejidad	Registros
131–135	0.0186	0.0890	0.1412	26.0147	23
136–140	0.0221	0.0868	0.1540	26.7076	18
141–145	0.0194	0.0778	0.1389	25.2481	10
146–150	0.0256	0.0993	0.1697	31.9148	13
151–155	0.0320	0.0785	0.1752	26.1264	8

Posteriormente, se aplicó la prueba de normalidad de Shapiro-Wilk sobre cada variable implicada. La Tabla 4.9 muestra que todos los valores p fueron significativamente menores a 0.05, indicando que ninguna de las métricas ni el "número de palabras" sigue una distribución normal. Esto justificó el uso de análisis no paramétricos y visualizaciones descriptivas.

Tabla 4.9. Prueba de normalidad con Shapiro-Wilk.

Métrica	Estadístico	P_valor	Interpretación
núm_palabras	0.83412	$1.98e-43$	Se rechaza la hipótesis nula: Los datos no siguen una distribución normal.
BLEU	0.74329	$2.11e-50$	Se rechaza la hipótesis nula: Los datos no siguen una distribución normal.
ROUGE_L	0.86558	$2.92e-40$	Se rechaza la hipótesis nula: Los datos no siguen una distribución normal.
$F_1 - Score$	0.94677	$8.52e-28$	Se rechaza la hipótesis nula: Los datos no siguen una distribución normal.
Perplejidad	0.02763	$5.36e-75$	Se rechaza la hipótesis nula: Los datos no siguen una distribución normal.

La Figura 4.3.3 muestra histogramas, curvas KDE y la campana de Gauss para cada métrica, permitiendo observar la asimetría y dispersión de los datos, lo que respalda el uso de métodos no paramétricos.

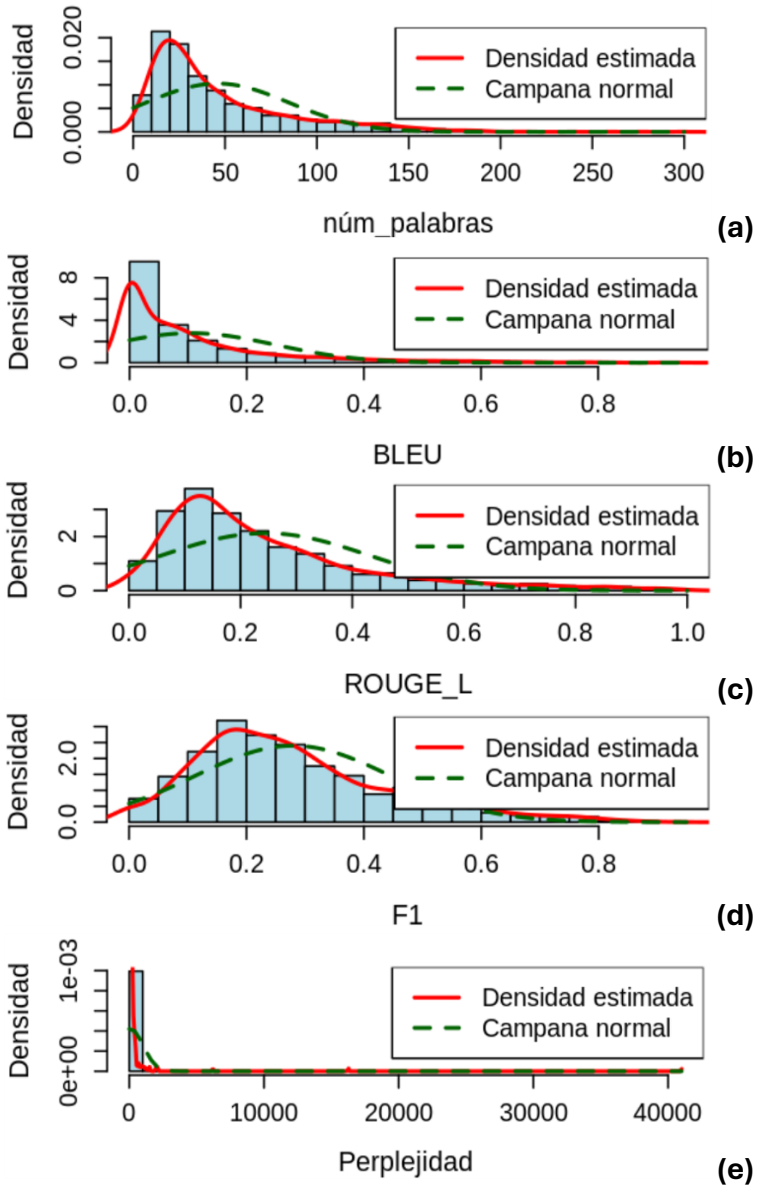

Figura 4.8: Distribución normal, histograma y prueba de densidad de las diferentes métricas.

Finalmente, los resultados mostrados en la Tabla 4.8 evidencian que los rangos de 11–15, 16–20 y 21–25 palabras concentran los valores más altos en las métricas BLEU, ROUGE_L y $F_1 - Score$, junto con una disminución progresiva en la Perplejidad. Esto sugiere que, dentro de dichos intervalos, las respuestas humanas evaluadas presentan una mayor coherencia semántica, precisión léxica y correspondencia entre pregunta y respuesta.

Con base en estos resultados, donde los intervalos de 21–25 palabras mostraron el mayor equilibrio entre alta puntuación y baja perplejidad, se definió —por decisión metodológica— un valor promedio de 25 palabras como tamaño óptimo de respuesta.

Para complementar el análisis, se realizó una visualización multivariada del "número de palabras" frente a las métricas BLEU, ROUGE_L y $F_1 - Score$ como se puede ver en la Figura 4.9, lo que facilitó la interpretación de relaciones y tendencias observadas.

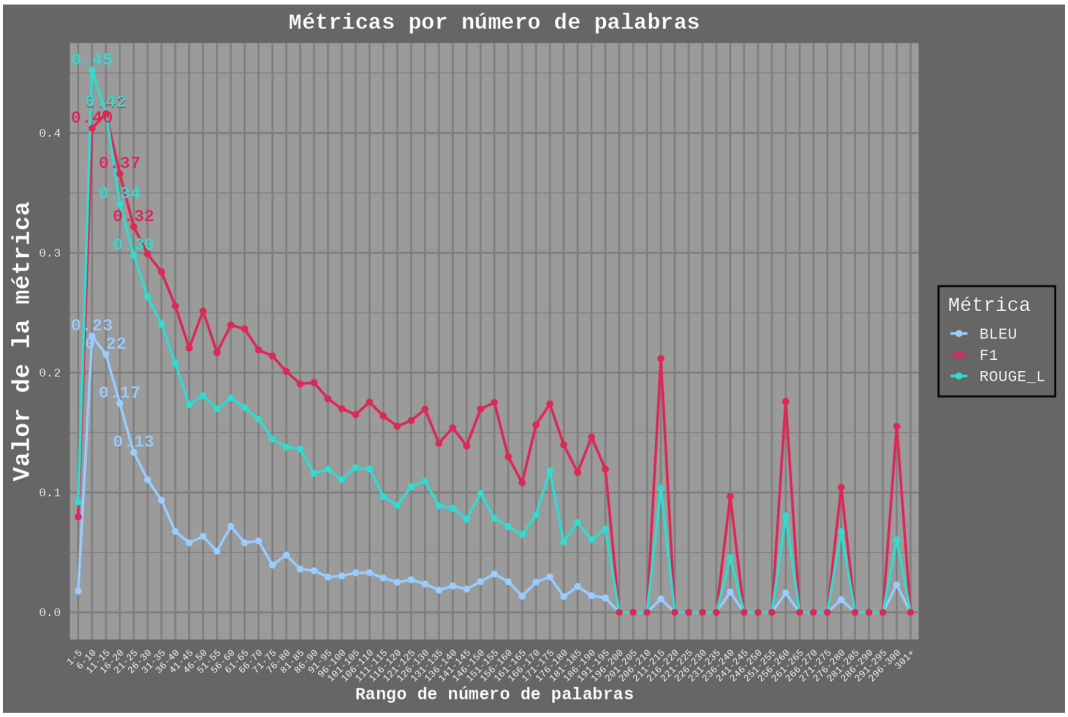

Figura 4.9: Visualización multivariada de "número de palabras" y métricas BLEU, ROUGE_L y $F_1 - Score$.

4.4 Métricas de Evaluación de *Prompts*

La evaluación de los *Prompts* constituye una etapa fundamental para analizar la calidad y efectividad de las interacciones con los LLMs. Mediante el uso de métricas cuantitativas y cualitativas es posible determinar el grado de coherencia, relevancia y precisión en las respuestas generadas por el LLM.

El presente apartado describe el conjunto de métricas empleadas para valorar el desempeño de los *Prompts* utilizados durante la fase experimental, así como su papel dentro del proceso de validación de resultados. Estas métricas permiten establecer una base objetiva para comparar configuraciones, medir mejoras y respaldar las conclusiones obtenidas en los resultados experimentales y en la validación de hipótesis posteriores.

4.4.1 Evaluar la Métrica de Calidad de *Prompts*

La evaluación de la calidad de los *Prompts* es crucial para optimizar la interacción con los LLMs.

Para este estudio, se emplearon métricas que permiten analizar aspectos específicos de los *Prompts*, centrándose en tres dimensiones fundamentales:

- **Coherencia**: Evalúa la consistencia interna del *Prompt*, asegurando que la formulación sea clara y lógica. Un *Prompt* coherente permite generar respuestas estructuradas y precisas.
- **Relevancia**: Mide la capacidad del *Prompt* para transmitir información significativa y útil al LLM, garantizando respuestas contextualmente apropiadas.
- **Diversidad**: Representa la variedad de respuestas posibles ante un mismo *Prompt*, evitando repeticiones y favoreciendo la adaptabilidad del LLM a distintos contextos.

Descripción del *dataset* y proceso de adquisición de datos

En coherencia con lo descrito en la Etapa 1: Recolección y Preparación de Datos, en una primera extracción se obtuvieron 883 *Prompts* de la comunidad de Reddit, específicamente del foro `/preguntaleareddit`. Esta fuente fue seleccionada por su diversidad temática y de estilo, lo que permitió capturar una amplia gama de estructuras lingüísticas y niveles de complejidad.

A partir de estos 883 *Prompts* originales, se generaron automáticamente 883 versiones tipo pregunta mediante el LLM GPT-4o, con el propósito de normalizar su estructura y mejorar su claridad lingüística. De esta manera, el conjunto total empleado en esta etapa estuvo conformado por 1,766 *Prompts*, distribuidos equitativamente entre versiones originales y versiones reformuladas.

Los *Prompts* se clasificaron en dos categorías, como se muestra en la Tabla 4.10, distinguiendo entre versiones originales (tipo 0) y versiones corregidas o refinadas (tipo 1).

Tabla 4.10. Ejemplos de clasificación de *Prompts* tipo 0 y tipo 1.

Oración original (*Prompt* tipo 0)	**Oración corregida (*Prompt* tipo 1)**
Hombres, ¿qué le quita lo atractivo a una mujer?	Hombres, ¿qué aspectos consideran que disminuyen la atracción hacia una mujer?
Aburrido, manden DM, quiero platicar con alguna chica.	Estoy aburrido, envíenme un mensaje directo; deseo platicar con alguna chica.

Implementación del Análisis de Componentes Principales

El análisis de componentes principales (PCA) se aplicó para explorar cómo las métricas de coherencia, relevancia, diversidad y la métrica propuesta, PromptIQ, se relacionan entre sí. En este caso, el PCA no se utilizó como una técnica de reducción de dimensionalidad —dado que el número de variables es limitado—, sino como un método estadístico para validar la capacidad descriptiva de PromptIQ frente a las métricas tradicionales y determinar su peso relativo en la explicación de la calidad global de los *Prompts*.

El PCA permite identificar qué variables concentran mayor varianza y, por tanto, mayor poder explicativo en la evaluación de calidad. Este enfoque facilita comprobar si la métrica propuesta captura de manera más completa la información que las métricas tradicionales.

La función `fviz_contrib()` de R se empleó para visualizar la contribución de cada métrica a las dimensiones principales del análisis:

- **Contribución de las Métricas:** Permiten observar gráficamente el peso de cada métrica dentro de las dimensiones principales.
- **Interpretación Visual:** Facilita identificar las métricas con mayor influencia en la varianza explicada.
- **Beneficio del Análisis:** Ayuda a enfocar la evaluación en las métricas que mejor representan la calidad general de los *Prompts*.

Resultados del análisis

Los resultados obtenidos del análisis de PCA muestran que la métrica PromptIQ presenta la mayor contribución en la primera componente principal (PC1), que explica el 60.29 % de la varianza total como puede observarse en la Tabla 4.11. Esto indica que PromptIQ concentra la mayor parte de la información relevante para evaluar la calidad de los *Prompts*.

Tabla 4.11. Desviación estándar, proporción de varianza y proporción acumulada por componente principal.

Importancia de los componentes principales	PC1	PC2	PC3	PC4
Desviación estándar	$\mathbf{0.142}$	0.081	0.065	0.051
Proporción de la varianza	$\mathbf{0.603}$	0.195	0.125	0.077
Proporción acumulada	$\mathbf{0.603}$	0.798	0.923	1.000

El resultado indica que PromptIQ posee la mayor carga factorial sobre el primer componente principal (PC1), con un valor de -0.985. Este coeficiente, aunque negativo, no implica una contribución desfavorable, sino que señala una relación inversa con la dirección del eje principal definido por el modelo. En el contexto del análisis de componentes principales, el signo

sólo refleja la orientación matemática del vector y no afecta la magnitud de la contribución.

Por tanto, su valor absoluto confirma que PromptIQ es la métrica dominante, explicando la mayor parte de la varianza asociada al PC1 como se muestra en la Tabla 4.12, actuando como el principal descriptor del comportamiento general de calidad de los *Prompts*.

Tabla 4.12. PromptIQ como métrica dominante en la descripción del componente principal PC1.

Métrica	PC1	PC2	PC3	PC4
PromptIQ	$\mathbf{-0.985}$	0.133	0.106	-0.009
Coherencia	-0.145	-0.933	-0.201	-0.259
Diversidad	0.027	0.298	-0.200	-0.933
Relevancia	0.085	-0.149	0.953	-0.250

La Figura 4.10 ilustra gráficamente las relaciones obtenidas entre las métricas en el análisis de PCA.

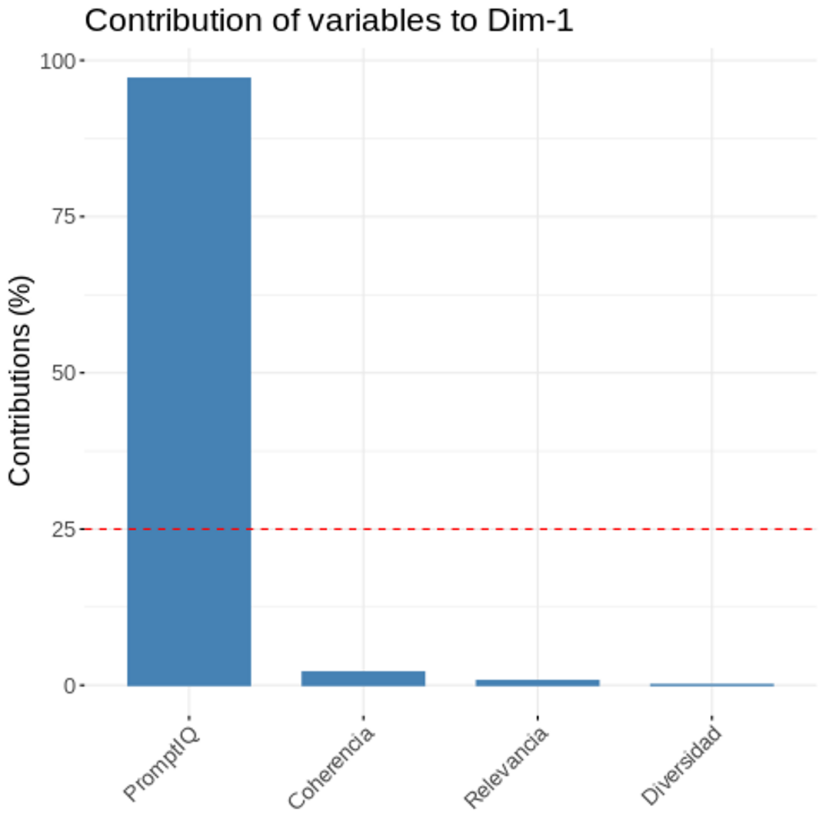

Figura 4.10: Evaluación de PromptIQ mediante análisis de PCA.

En conclusión, los resultados indican que PromptIQ es la métrica más representativa para evaluar la calidad de los *Prompts*, al capturar de forma más integral los aspectos de coheren-

cia, relevancia y diversidad.

Aunque las métricas tradicionales siguen siendo útiles, su contribución relativa es menor, reafirmando la utilidad de PromptIQ como referencia principal para la evaluación y optimización de *Prompts* en LLMs.

4.5 Resultados Experimentales

En esta sección se presentan los principales hallazgos obtenidos durante la evaluación experimental de la plataforma web desarrollada para la mejora automatizada de *Prompts*. Los resultados exponen tanto el desempeño funcional del sistema en diversos casos de uso —considerando distintos niveles de complejidad y tipos de errores lingüísticos— como el impacto de la calidad de los *Prompts* en el tiempo de respuesta de un LLM.

Se incluyen pruebas orientadas a validar la efectividad del sistema para detectar, corregir y optimizar entradas textuales, así como análisis estadísticos que permiten comprender la relación entre la calidad del *Prompt* y el comportamiento del LLM generativo. En conjunto, estos resultados proporcionan evidencia empírica sobre la utilidad, precisión y eficiencia de la solución propuesta.

4.5.1 Crear Interfaz Web de Consultas

En esta sección se describe el desarrollo de una interfaz web orientada a la optimización automática de *Prompts*, diseñada para usuarios sin conocimientos técnicos en LLMs. Esta herramienta constituye la fase práctica de implementación descrita en la Sección 3.2.1, en la cual se definió el conjunto de datos base para entrenar y validar el sistema.

La plataforma implementa la métrica PromptIQ, descrita en el Capítulo 4.4.1, como criterio principal de evaluación de la calidad de los *Prompts*. A través de una interfaz intuitiva y un motor basado en GPT-4o, el sistema analiza la entrada del usuario, detecta errores y genera versiones mejoradas con una puntuación PromptIQ superior, facilitando así la redacción de consultas claras y efectivas.

Para evaluar la funcionalidad de la interfaz, se diseñaron distintos casos de prueba que abarcan escenarios reales de interacción por parte de usuarios no especializados. Estos casos permiten validar la precisión del sistema, la efectividad de las sugerencias generadas y la estabilidad del flujo de interacción.

Caso de prueba CP-01.

Elemento	Descripción
ID del caso de prueba	CP-01
Nombre	Generación y validación de Prompts a partir de una palabra seguida de signos de puntuación.
Objetivo	Verificar que el sistema detecte errores lingüísticos en Prompts mal escritos y genere sugerencias bien estructuradas.
Entrada	Prompt inicial: “mexico , , , , , , , , , ”.
Procedimiento	1. Ingresar el Prompt con errores de puntuación. 2. Esperar a que el sistema genere tres sugerencias tipo pregunta. 3. Seleccionar una sugerencia. 4. Verificar que la respuesta sea coherente. 5. Confirmar la aparición del botón azul para continuar.
Resultado esperado	1. El sistema detecta múltiples errores. 2. Se generan 3 sugerencias sin errores lingüísticos. 3. La respuesta es relevante y no excede 25 palabras. 4. Se habilita el botón azul de continuar.
Figuras relacionadas	Véase la Figura 4.11.
Observaciones	1. El Prompt original obtuvo 0.3463. 2. Las tres sugerencias obtuvieron PromptIQ 0.9987. 3. La respuesta fue correcta. 4. El botón azul apareció correctamente.

Figura 4.11: Interfaz web del caso de prueba CP-01

Caso de prueba CP-02.

Elemento	Descripción
ID del caso de prueba	CP-02
Nombre	Evaluación de un Prompt correctamente formulado sobre historia urbana de la CDMX.
Objetivo	Verificar que el sistema reconozca un Prompt bien estructurado, sin errores lingüísticos, y calcule correctamente su puntuación de calidad (PromptIQ), sin generar sugerencias innecesarias.
Entrada	*Prompt*: "¿Qué papel desempeñó la Calle José María Pino Suárez en la planificación urbana de la Ciudad de México?"
Procedimiento	1. Ingresar el Prompt directamente en el campo de entrada. 2. Observar si el sistema identifica errores lingüísticos. 3. Verificar la puntuación PromptIQ asignada. 4. Confirmar si el sistema sugiere mejoras o permite continuar directamente. 5. Verificar que se habilite el botón de "continuar" o una respuesta si el Prompt es aceptado.

Caso de prueba CP-02 (*continuación*).

Elemento	Descripción
Resultado esperado	1. El sistema debe reconocer que el Prompt no contiene errores. 2. La puntuación PromptIQ debe ser alta (mayor a 0.95). 3. No deben generarse sugerencias de mejora. 4. El sistema debe permitir al usuario continuar con el flujo o recibir una respuesta directa.
Figura relacionada	Véase la Figura 4.12.
Observaciones	1. El Prompt ingresado obtuvo una puntuación PromptIQ de 0.9987. 2. No se detectaron errores en ninguna de las categorías evaluadas. 3. El sistema no generó sugerencias, indicando que el Prompt fue aceptado como óptimo. 4. El flujo permitió continuar sin interrupciones o correcciones adicionales.

Figura 4.12: Interfaz y resultado del caso de prueba CP-02.

Caso de prueba CP-03.

Elemento	Descripción
ID del caso de prueba	CP-03

Caso de prueba CP-03 (*continuación*).

Elemento	Descripción
Nombre	Corrección y mejora de un Prompt incompleto con dos palabras y comas aisladas.
Objetivo	Evaluar la capacidad del sistema para detectar errores en un Prompt mal formado y generar automáticamente tres sugerencias coherentes y bien estructuradas, con su respectiva puntuación PromptIQ.
Entrada	*Prompt*: “pino , , , , , suarez”
Procedimiento	1. Ingresar el Prompt mal estructurado en el sistema. 2. Observar la identificación de errores lingüísticos. 3. Analizar las tres sugerencias generadas por el sistema. 4. Comparar la puntuación PromptIQ del Prompt original y de los sugeridos. 5. Evaluar si las sugerencias son coherentes con el contexto esperado (historia urbana de CDMX). 6. Verificar si el sistema permite continuar el flujo tras seleccionar una sugerencia.
Resultado esperado	1. El sistema debe detectar múltiples errores en el Prompt original. 2. Se deben generar tres nuevos Prompts bien redactados y relevantes. 3. Los nuevos Prompts deben alcanzar una puntuación PromptIQ mayor a 0.93. 4. El usuario debe poder seleccionar un Prompt sugerido y continuar con la interacción sin errores.
Figura relacionada	Véase la Figura 4.13.
Observaciones	1. El Prompt original obtuvo una puntuación PromptIQ de 0.749, con errores de puntuación, ortografía, sintaxis y fluidez. 2. El sistema generó tres sugerencias con valores PromptIQ de 0.9987, 0.9807 y 0.9807. 3. El primer Prompt sugerido fue: “¿Cuál es la historia detrás del nombre de la estación de metro Pino Suárez en la Ciudad de México?” 4. Las sugerencias fueron coherentes con el dominio histórico y urbano. 5. El sistema permitió seleccionar un Prompt corregido y continuar con el flujo sin errores.

Figura 4.13: Interfaz y resultado del caso de prueba CP-03.

Caso de prueba CP-04.

Elemento	Descripción
ID del caso de prueba	CP-04
Nombre	Corrección de un Prompt informal con emojis y falta de estructura lingüística.
Objetivo	Verificar si el sistema es capaz de identificar un Prompt informal con uso de emojis y falta de sintaxis clara, y generar sugerencias bien estructuradas, coherentes y relevantes al contexto cultural del Zócalo de la Ciudad de México.
Entrada	*Prompt*: ":) zocalo mexicano ?????"
Procedimiento	1. Ingresar el Prompt informal en la interfaz del sistema. 2. Registrar los errores lingüísticos detectados en el Prompt original. 3. Analizar las tres sugerencias de Prompts generadas automáticamente. 4. Comparar las puntuaciones PromptIQ entre el Prompt original y las sugerencias. 5. Confirmar que el sistema permite seleccionar una sugerencia para continuar con la generación de respuesta.

Caso de prueba CP-04 (*continuación*).

Elemento	Descripción
Resultado esperado	1. El sistema debe identificar múltiples errores lingüísticos en el Prompt original. 2. Se deben generar tres Prompts tipo pregunta correctamente estructurados. 3. Las nuevas sugerencias deben tener una puntuación PromptIQ ≥ 0.93. 4. La temática de los Prompts sugeridos debe estar alineada con eventos culturales en el Zócalo de la Ciudad de México.
Figura relacionada	Véase la Figura 4.14.
Observaciones	1. El Prompt original obtuvo una puntuación PromptIQ de 0.4227, con errores en puntuación, ortografía, fluidez, incoherencia, semántica y pragmatismo. 2. Se generaron tres sugerencias con PromptIQ de 0.9987 cada una, sin errores lingüísticos. 3. El sistema permitió seleccionar una de las opciones para continuar, cumpliendo con el flujo esperado de la interfaz.

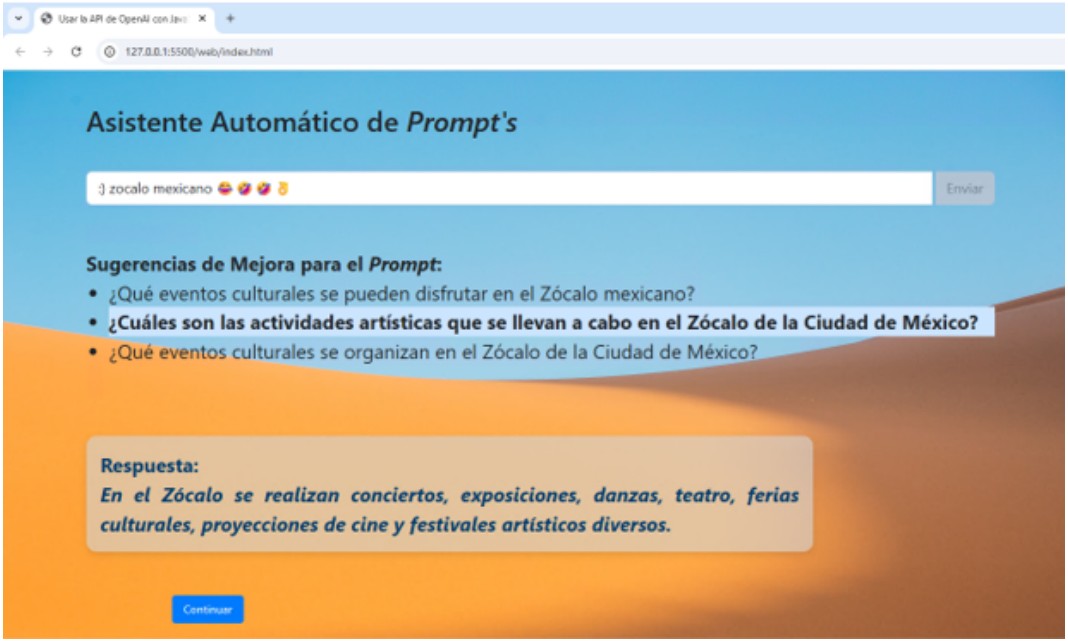

Figura 4.14: Interfaz y resultado del caso de prueba CP-04.

Los resultados demuestran que la interfaz logró mejorar de manera consistente la calidad de los *Prompts*, alcanzando valores de PromptIQ superiores a 0.98 en todos los casos con erro-

res de entrada. Esto evidencia la efectividad del sistema para guiar al usuario hacia la redacción óptima de consultas compatibles con LLMs.

4.5.2 Implementación y Experimentación

La plataforma web desarrollada para la mejora automatizada de *Prompts*, orientada a usuarios no expertos, demostró una alta eficacia en la detección y corrección de errores lingüísticos en diversos escenarios de uso. Durante las pruebas experimentales, el sistema identificó con precisión problemas de puntuación, sintaxis, ortografía, coherencia y pragmática, incluso en entradas informales o incompletas. Además, generó sugerencias gramaticalmente correctas y culturalmente adecuadas, permitiendo al usuario avanzar sin requerir conocimientos técnicos previos. La interfaz resultó intuitiva y eficiente, manteniendo un flujo de interacción natural y validando su usabilidad para un público amplio.

Los resultados evidenciaron que los *Prompts* mejorados por la plataforma alcanzaron puntuaciones PromptIQ superiores a 0.98, garantizando la calidad y pertinencia de las correcciones propuestas. En contraste, los *Prompts* originales, especialmente aquellos con errores evidentes o con lenguaje informal, presentaron valores significativamente más bajos, lo que confirma la necesidad de aplicar mecanismos automáticos de mejora lingüística.

La métrica PromptIQ fue esencial para evaluar la calidad lingüística antes y después del proceso de mejora. En los casos analizados, los *Prompts* originales registraron valores entre 0.34 y 0.75, reflejando deficiencias en puntuación, ortografía, fluidez y coherencia semántica. Tras el procesamiento automático, las tres sugerencias generadas por el sistema obtuvieron valores iguales o superiores a 0.98, alcanzando en algunos casos hasta 0.9987. Este incremento cuantitativo demuestra la efectividad del sistema para optimizar la estructura, claridad y relevancia temática, asegurando una generación de respuestas más precisas y coherentes.

Asimismo, la técnica de mejora implementada mostró un desempeño robusto y versátil, capaz de adaptarse a diferentes tipos de errores, desde fallas gramaticales hasta el uso inadecuado de emojis o expresiones coloquiales. La generación simultánea de tres sugerencias tipo pregunta por cada entrada, contextualizadas cultural e históricamente, facilitó la elección de opciones más claras y consistentes.

En conjunto, los resultados obtenidos confirman que la técnica de mejora automatizada no solo incrementa la calidad lingüística de los *Prompts*, sino que también optimiza el rendimiento general del sistema de generación, consolidándose como una herramienta eficaz para usuarios sin formación especializada.

4.5.3 Evaluar el Impacto en el Tiempo de Respuesta del LLM

La calidad del *Prompt* constituye un factor determinante en el desempeño de los LLMs, ya que influye directamente en la velocidad y precisión con que estos procesan la información.

Medir el tiempo de respuesta resulta esencial para evaluar la eficiencia operativa de los LLMs, especialmente en contextos donde la inmediatez es clave, como en sistemas conversacionales, asistentes virtuales o generación automática de contenido en tiempo real. Un *Prompt* bien estructurado no solo favorece respuestas más precisas, sino también una ejecución más rápida, optimizando recursos computacionales y mejorando la experiencia del usuario.

En coherencia con lo descrito en la Etapa 1: Recolección y Preparación de Datos, se emplearon 1,796 *Prompts* extraídos del foro `/preguntaleareddit`. Posteriormente, cada uno de estos *Prompts* fue reformulado en su versión tipo pregunta mediante el uso del LLM GPT-4o, con el propósito de estandarizar su estructura y claridad lingüística. Como resultado, el conjunto total analizado estuvo conformado por 3,592 *Prompts*, distribuidos equitativamente entre los originales y sus correspondientes versiones reformuladas.

En este análisis se buscó comprender cómo la formulación de los *Prompts* afecta el tiempo de respuesta del LLM. Para ello, los 3,592 *Prompts* se clasificaron en dos tipos: los originales extraídos de Reddit, considerados “malos” (tipo 0) por su baja calidad lingüística, con errores ortográficos, gramaticales o estructuras ambigüas; y sus correspondientes versiones reformuladas mediante GPT-4o, consideradas “buenos” (tipo 1) por ser claras, coherentes y gramaticalmente correctas. Las métricas evaluadas incluyeron Perplejidad, Gramaticalidad, coherencia y tiempo de respuesta (en segundos).

Antes de comparar estadísticamente los dos grupos, se aplicó la prueba de normalidad de Shapiro–Wilk *Test* para determinar si los datos seguían una distribución normal. Esta prueba es apropiada para tamaños de muestra moderados y permite identificar desviaciones significativas respecto a la normalidad. Dado que los valores p obtenidos fueron menores que 0.05 en todas las métricas, se rechazó la hipótesis de normalidad. Por ello, se optó por emplear la prueba no paramétrica de Wilcoxon, que no asume normalidad y permite comparar distribuciones entre dos grupos relacionados o independientes.

En la Tabla 4.13 se presentan los resultados de ambas pruebas, mostrando que solo la métrica “Tiempo en segundos” presentó una diferencia estadísticamente significativa ($p = 1.54 \times 10^{-14}$), mientras que las métricas Perplejidad, Gramaticalidad y coherencia no evidenciaron diferencias relevantes entre los dos tipos de *Prompts*.

Tabla 4.13. Resultados estadísticos de las diferentes métricas evaluadas para detectar cuál de ellas genera una diferencia significativa entre los datos analizados.

Métrica	Shapiro-Wilk		Prueba de Wilcoxon
	Prompts tipo 0	Prompts tipo 1	
Perplejidad	$W = 0.5272$, $p = 3.40 \times 10^{-56}$ (se rechaza H_0)	$W = 0.5375$, $p = 8.17 \times 10^{-56}$ (se rechaza H_0)	$U = 794355.5$, $p = 0.6192$ (no se rechaza H_0)
Gramaticalidad	$W = 0.9820$, $p = 3.89 \times 10^{-14}$ (se rechaza H_0)	$W = 0.9900$, $p = 1.09 \times 10^{-9}$ (se rechaza H_0)	$U = 700617$, $p = 0.5060$ (no se rechaza H_0)
Coherencia	$W = 0.8375$, $p = 3.34 \times 10^{-39}$ (se rechaza H_0)	$W = 0.8399$, $p = 5.55 \times 10^{-39}$ (se rechaza H_0)	$U = 756504.5$, $p = 0.3679$ (no se rechaza H_0)
Tiempo en segundos	$W = 0.3765$, $p = 4.08 \times 10^{-61}$ (se rechaza H_0)	$W = 0.3826$, $p = 6.18 \times 10^{-61}$ (se rechaza H_0)	$\boldsymbol{U = 593348.5}$, $\boldsymbol{p = 1.54 \times 10^{-14}}$ **(se rechaza H_0)**

Leyendas:

- W: Estadístico de Shapiro-Wilk.
- p: Valor p asociado.
- U: Estadístico de Wilcoxon.
- H_0: Hipótesis nula.

Los resultados de la prueba de Wilcoxon muestran diferencias estadísticamente significativas entre los dos tipos de *Prompts* únicamente en la métrica de "tiempo de respuesta", indicando que la velocidad de generación varía según la formulación del *Prompt*. En contraste, para las métricas de Perplejidad, Gramaticalidad y Coherencia no se encontraron diferencias significativas, lo que sugiere que la calidad lingüística del texto generado se mantiene relativamente estable entre ambos tipos de *Prompts*.

En conclusión, este experimento demuestra que mejorar la claridad y estructura de los *Prompts* tiene un efecto directo sobre la eficiencia temporal de los LLMs. Si bien no se evidenciaron mejoras sustanciales en las métricas lingüísticas, la reducción en el tiempo de respuesta representa un beneficio práctico significativo para aplicaciones que demandan rapidez, como asistentes conversacionales o sistemas de generación automática de contenido en tiempo real.

4.5.4 Validación de Hipótesis

La presente sección se centra en la validación de la hipótesis planteada en esta investigación que: ***"La aplicación de Ingeniería de Prompts mejora la calidad de los Prompts redactados por usuarios no expertos"***. Para ello, se analizarán los resultados obtenidos a partir de la implementación de la métrica propuesta PromptIQ.

De esta forma, la sección ofrece un marco sistemático para corroborar la hipótesis central y evaluar el impacto de la Ingeniería de *Prompts* en la mejora de la interacción con los LLMs, proporcionando bases sólidas para los análisis y conclusiones finales del estudio.

4.5.5 Análisis del Impacto de los *Prompts* Mejorados

La aplicación de la metodología desarrollada para la mejora automática de *Prompts* mediante Ingeniería de *Prompts* ha permitido evaluar de manera sistemática cómo las reformulaciones impactan la calidad y eficiencia de las interacciones con los LLMs.

El uso de un modelo de regresión lineal multivariable basado en el conteo de errores lingüísticos facilitó cuantificar la calidad de los *Prompts* escritos por usuarios no expertos, mediante la métrica propuesta en este trabajo PromptIQ. Los resultados muestran que la técnica automática de transformación de *Prompts* mal estructurados hacia un formato tipo pregunta incrementa significativamente su claridad, coherencia y estructura, elevando consistentemente su valor de PromptIQ por encima del umbral de 0.84.

La integración del LLM GPT-4o en la plataforma automatizada permitió generar sugerencias mejoradas de manera escalable, robusta y accesible, demostrando que la técnica es efectiva tanto para optimizar la calidad lingüística de los *Prompts* como para agilizar la interacción con los LLMs.

Asimismo, se observó que los *Prompts* mejorados no solo incrementan la calidad del contenido generado, sino que también generan respuestas más rápidas en GPT-4o, mostrando un impacto positivo en la eficiencia operativa del LLM. En conjunto, estos hallazgos corroboran que la Ingeniería de *Prompts* mejora sustancialmente la calidad de los *Prompts* producidos por usuarios no expertos, validando la hipótesis central y evidenciando el potencial de estas técnicas para optimizar la interacción con los LLMs.

Capítulo 5

Conclusiones

A continuación, se presentan las conclusiones y las proyecciones a futuro derivadas de esta investigación. Se resumen los principales aportes metodológicos y técnicos en la mejora de *Prompts* en español culto, así como el desarrollo de la métrica PromptIQ. Finalmente, se proponen líneas de investigación que amplían el alcance y aplicabilidad del trabajo en otros contextos lingüísticos y temáticos.

5.1 Conclusiones Generales

La presente investigación permitió desarrollar una técnica sistematizada para la mejora de *Prompts* en español culto, utilizando principios de Ingeniería de *Prompts* y NLP. Mediante el uso del LLM GPT-4o y estrategias como el FT junto con el análisis de métricas textuales, se estableció una metodología capaz de optimizar la calidad de las entradas dirigidas a un LLM.

Además, se propuso la métrica PromptIQ como herramienta para evaluar la calidad de los *Prompts*, diferenciando entre entradas deficientes y optimizadas. Se determinaron umbrales clave, incluyendo el número mínimo y máximo de palabras, que permiten clasificar un *Prompt* como claro, coherente y eficaz.

Esta investigación proporciona un marco práctico y evaluable para mejorar la interacción humano-modelo, promoviendo un uso más eficiente y preciso de los LLMs en español. Los resultados sientan bases para futuras investigaciones en dominios específicos y evidencian el potencial de combinar estrategias computacionales con criterios lingüísticos para optimizar la generación de texto automatizado.

Uno de los principales aportes fue demostrar que es posible automatizar la mejora de *Prompts* escritos por usuarios no especializados, facilitando su interacción con LLMs. Mediante un enfoque que combina criterios lingüísticos y estrategias de NLP, se logró transformar entradas deficientes en *Prompts* claros, coherentes y útiles. La implementación de PromptIQ permitió evaluar objetivamente la calidad de cada entrada y establecer umbrales mínimos de claridad

y extensión de las respuestas generadas por un LLM. Gracias a este proceso, el sistema no solo detecta errores comunes, sino que también produce sugerencias optimizadas que respetan la intención original del usuario.

En conjunto, los hallazgos validan que la combinación de técnicas de Ingeniería de *Prompts*, análisis estadístico y NLP puede mejorar significativamente la interacción con los LLMs, estableciendo un referente práctico para futuras aplicaciones y estudios en NLP aplicados al español.

5.2 Objetivos Alcanzados

La investigación alcanzó de manera satisfactoria sus objetivos principales, orientados al diseño, implementación y validación de una metodología para la mejora de *Prompts* en español culto. A continuación, se detallan los logros más relevantes:

- Se diseñó y sistematizó una técnica para la reescritura optimizada de *Prompts*, basada en principios de Ingeniería de *Prompts* y NLP.
- Se aplicó el LLM GPT-4o en tareas de mejora automática de *Prompts*, mediante estrategias como FT.
- Se definieron parámetros cuantificables sobre la estructura ideal de un *Prompt*, incluyendo umbrales mínimos y máximos de longitud y criterios de claridad lingüística.
- Se demostró que es posible mejorar automáticamente *Prompts* deficientes sin intervención humana experta, facilitando así el acceso a los LLMs.

Estos objetivos establecen las bases de un marco práctico, replicable y aplicable para mejorar la interacción entre usuarios y LLMs en contextos hispanohablantes.

5.3 Productos Generados

A partir de los resultados de esta investigación, se generaron diversos productos con valor metodológico, técnico y aplicado:

- **Metodología sistematizada** para la mejora de *Prompts* en español culto, orientada a usuarios no expertos.
- **Métrica PromptIQ**, desarrollada como instrumento de evaluación objetiva de la calidad de los *Prompts*, considerando claridad, extensión y coherencia semántica.
- **Prototipo funcional** de sistema automatizado de mejora de *Prompts*, basado en el LLM GPT-4o.

- **Plataforma web interactiva** para la mejora automatizada de *Prompts*, dirigida a usuarios sin conocimientos técnicos, que incluye análisis, detección de errores y generación de sugerencias optimizadas con interfaz accesible.
- **Informe de buenas prácticas** para el diseño de *Prompts* eficaces en español, adaptable a otros dominios y contextos lingüísticos.

Estos productos representan un aporte significativo a la Ingeniería de *Prompts* en español, con aplicaciones potenciales en educación, tecnologías del lenguaje y desarrollo de sistemas conversacionales, así como una base replicable para futuras investigaciones.

5.4 Trabajo Futuro

Como extensión de esta investigación, se identifican diversas líneas de trabajo futuro. Entre ellas destaca la validación de la métrica PromptIQ en otros contextos lingüísticos, dominios temáticos y variantes del español, con el fin de evaluar su capacidad de generalización y robustez. Asimismo, resulta pertinente explorar la integración de técnicas de aprendizaje activo para que el sistema de mejora de *Prompts* pueda adaptarse dinámicamente al estilo de escritura del usuario, optimizando su desempeño en tiempo real.

Estas propuestas buscan fortalecer la interacción entre usuarios humanos y LLMs, haciendo más accesible, eficiente y productivo el uso de esta tecnología en contextos reales y aplicados.

Referencias

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. *Information Systems*, *112*, 102131.

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

Amur, Z. H., Kwang Hooi, Y., Bhanbhro, H., Dahri, K., & Soomro, G. M. (2023). Short-text semantic similarity (stss): Techniques, challenges and future perspectives. *Applied Sciences*, *13*(6), 3911.

Arqueología Mexicana. (2024). CDMX Guía de Viajeros. Parte 1. Edición Especial 115. https://arqueologiamexicana.mx/

Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: a review. *Artificial intelligence review*, *56*(9), 10345-10425.

Bakar, Z. A., Ismail, N. K., & Rawi, M. I. M. (2017). Identification of Noun+ Verb Compound Nouns in Malay Standard document based on rule based. *2017 IEEE 3rd International Conference on Engineering Technologies and Social Sciences (ICETSS)*, 1-6.

Bergmann, D. (2024, marzo). *What is Fine-Tuning?* IBM. https://www.ibm.com/think/topics/fine-tuning

Borden, J. H., Mahajan, U. V., Eyasu, L., Holden, W., Shaw, B., Callas, P., & Benzil, D. L. (2022). Evaluating diversity in neurosurgery through the use of a multidimensional statistical model: a pilot study. *Journal of Neurosurgery*, *137*(3), 859-866.

Burlin, C. (2023). Explainability to enhance creativity: A human-centered approach to prompt engineering and task allocation in text-to-image models for design purposes [diVA portal, Preprint].

Candelo, J. E., Soto, J. D., Torres, L., Schettini, N., Calle, M., Garcia, L., & de Castro, A. (2018). Coherence and cohesion issues in argumentation documents written by engineering students. *2018 IEEE Global Engineering Education Conference (EDUCON)*, 156-160.

Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity—a survey. *Acm Computing Surveys (Csur)*, *54*(2), 1-37.

Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2023). A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM transactions on software engineering and methodology*, *32*(4), 1-30.

Churchill, R., & Singh, L. (2022). The evolution of topic modeling. *ACM Computing Surveys*, *54*(10s), 1-35.

Cuberos, R. (2019). *Indicadores léxicos de calidad textual en español nativo y no nativo* [Tesis doctoral, Universitat de Barcelona].

DAIR.AI. (2025). *Guía de ingeniería de prompts*. Prompt Engineering Guide. https://www.promptingguide.ai/es

Denny, P., Leinonen, J., Prather, J., Luxton-Reilly, A., Amarouche, T., Becker, B. A., & Reeves, B. N. (2023). Promptly: Using prompt problems to teach learners how to effectively utilize AI code generators [Preprint]. *arXiv preprint arXiv:2307.16364*.

Diccionario de la lengua española. (2024). Técnica. https://dle.rae.es/t%C3%A9cnico

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., & Goldberg, Y. (2021). Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, *9*, 1012-1031.

Ganesan, K. (2018). Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *arXiv preprint arXiv:1803.01937*.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, *36*(2), 193-202.

Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W., & Gao, C. (2021). A survey on the techniques, applications, and performance of short text semantic similarity. *Concurrency and Computation: Practice and Experience*, *33*(5), e5971.

Huang, H., Zhang, B., Li, Y., Zhang, B., Sun, Y., Luo, C., & Peng, C. (2023). Knowledge-enhanced prompt-tuning for stance detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(6), 1-20.

Hugging Face. (s.f.). Hugging Face Model Hub [Consultado en noviembre de 2025]. https://huggingface.co/models

Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). How can we know what language models know? *Transactions of the Association for Computational Linguistics*, *8*, 423-438.

Jin, F., Lu, J., Zhang, J., & Zong, C. (2023). Instance-aware prompt learning for language understanding and generation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(7), 1-18.

Li, L., Zhang, Y., & Chen, L. (2023). Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, *41*(4), 1-26.

Li, X., Peng, D., & Wang, Y. (2022). Improving patient self-description in Chinese online consultation using contextual prompts. *BMC Medical Informatics and Decision Making*, *22*(1), 170.

Liddy, E. D. (2001). *Natural Language Processing*. Marcel Dekker, Inc. https://surface.syr.edu/istpub

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, *55*(9), 1-35.

Martín, A., Huertas-Tato, J., Huertas-García, Á., Villar-Rodríguez, G., & Camacho, D. (2022). FacTeR-Check: Semi-automated fact-checking through semantic similarity and natural language inference. *Knowledge-based systems*, *251*, 109265.

McCoy, R. T., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, *11*, 652-670.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. *Computers & Education*, *56*(2), 337-345.

OpenAI. (s.f.). OpenAI API Documentation [Consultado en noviembre de 2025]. https://platform.openai.com/docs/api-reference

Park, D., An, G. T., Kamyod, C., & Kim, C. G. (2023). A study on performance improvement of prompt engineering for generative AI with a large language model. *Journal of Web Engineering*, *22*(8), 1187-1206.

Popovic, V. (2000). Expert and novice user differences and implications for product design and useability. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(38), 933-936.

Ramos, J., et al. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*, *242*(1), 29-48.

Rasheed Issam, K. A., Patel, S., et al. (2021). Topic Modeling Based Extractive Text Summarization. *arXiv e-prints*, arXiv-2106.

Rosario, G., & Noever, D. (2023). Grading conversational responses of chatbots. *arXiv preprint arXiv:2303.12038*.

Shahi, T. B., & Sitaula, C. (2022). Natural language processing for Nepali text: A review. *Artificial Intelligence Review*, *55*(4), 3401-3429.

Singh, C., Morris, J. X., Aneja, J., Rush, A. M., & Gao, J. (2023). Explaining data patterns in natural language with language models. *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 31-55.

Sitarz, M. (2022). Extending F1 metric, probabilistic approach. *arXiv preprint arXiv:2210.11997*.

Strobelt, H., Webson, A., Sanh, V., Hoover, B., Beyer, J., Pfister, H., & Rush, A. M. (2022). Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE transactions on visualization and computer graphics*, *29*(1), 1146-1156.

Tatbul, N., Lee, T. J., Zdonik, S., Alam, M., & Gottschlich, J. (2018). Precision and recall for time series. *Advances in neural information processing systems*, *31*.

Terragni, S., Fersini, E., Galuzzi, B. G., Tropeano, P., & Candelieri, A. (2021). OCTIS: Comparing and optimizing topic models is simple! *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 263-270.

Thomason, J., Padmakumar, A., Sinapov, J., Walker, N., Jiang, Y., Yedidsion, H., Hart, J., Stone, P., & Mooney, R. (2020). Jointly improving parsing and perception for natural language commands through human-robot dialog. *Journal of Artificial Intelligence Research*, *67*, 327-374.

Thorne, J., Yazdani, M., Saeidi, M., Silvestri, F., Riedel, S., & Halevy, A. (2021). From natural language processing to neural databases. *Proceedings of the VLDB Endowment*, *14*(6), 1033-1039.

Velásquez-Henao, J. D., Franco-Cardona, C. J., & Cadavid-Higuita, L. (2023a). Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering. *Dyna*, *90*(SPE230), 9-17.

Velásquez-Henao, J. D., Franco-Cardona, C. J., & Cadavid-Higuita, L. (2023b). Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering. *Dyna*, *90*(SPE230), 9-17.

Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., & Zhang, S. (2023). Prompt engineering for healthcare: Methodologies and applications [Preprint]. *arXiv preprint arXiv:2304.14670*.

Wong, M., Ong, Y.-S., Gupta, A., Bali, K. K., & Chen, C. (2023). Prompt evolution for generative ai: A classifier-guided approach. *2023 IEEE Conference on Artificial Intelligence (CAI)*, 226-229.

Xu, S., Pang, L., Shen, H., & Cheng, X. (2023). Nir-prompt: A multi-task generalized neural information retrieval training framework. *ACM Transactions on Information Systems*, *42*(2), 1-32.

Yenduri, G., Srivastava, G., Maddikunta, P. K. R., Jhaveri, R. H., Wang, W., Vasilakos, A. V., Gadekallu, T. R., et al. (2023). Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *arXiv preprint arXiv:2305.10435*.

Yue, Z., Rong, C., Wang, Y., & Yang, Y. (2015). Lip-reading based on fuzzy language model. *2015 International Conference on Wireless Communications & Signal Processing (WCSP)*, 1-5.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-21.

Zhang, C., Zheng, J., Wang, Z., Jia, Z., & Li, F. (2018). A New Evaluation Criterion with the Integration of Perplexity and Jensen-Shannon Divergence for Biterm Topic Model. *2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, 283-287.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, *1*(2).

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision*, *130*(9), 2337-2348.

Zhou, Y.-C., Zheng, Z., Lin, J.-R., & Lu, X.-Z. (2022). Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking. *Computers in Industry*, *142*, 103746.

Anexo A: Constancias de Ponencias.

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

EL TECNOLÓGICO NACIONAL DE MÉXICO
A TRAVÉS DEL INSTITUTO TECNOLÓGICO DE ZACATEPEC
OTORGA EL PRESENTE

RECONOCIMIENTO

A

CARLOS JAVIER MARTÍNEZ JOFFROY

POR SU DESTACADA PARTICIPACIÓN CON LA CONFERENCIA TITULADA:

"TÉCNICAS DE SIMILITUD EN PROCESAMIENTO DE LENGUAJE NATURAL"

EN EL MARCO DEL

CONGRESO NACIONAL DE INNOVACIÓN, TECNOLOGÍA, LIDERAZGO Y SOSTENIBILIDAD 2024

ZACATEPEC, MOR., A 24 DE ABRIL DE 2024

LORENZO O. HERNÁNDEZ
DIRECTOR

EDUCACIÓN
INSTITUTO TECNOLÓGICO DE ZACATEPEC
DIRECCIÓN

100% PLÁSTICO

Anexo A.1: Ponencia titulada: "Técnicas de Similitud en Procesamiento de Lenguaje Natural".

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

EL TECNOLÓGICO NACIONAL DE MÉXICO
A TRAVÉS DEL INSTITUTO TECNOLÓGICO DE ZACATEPEC

OTORGA EL PRESENTE

RECONOCIMIENTO

A

CARLOS JAVIER MARTÍNEZ JOFFROY

POR SU DESTACADA PARTICIPACIÓN COMO PONENTE DE LA CONFERENCIA "FUTURO DIGITAL: EXPLORANDO MEDIAPIPE Y YOLO.", REALIZADO DEL 23 Y 24 DE OCTUBRE 2024 EN LAS JORNADAS ACADÉMICAS DE INNOVACIÓN, TECNOLOGÍA, LIDERAZGO Y SOSTENIBILIDAD 2024

ZACATEPEC, MOR., A 24 DE OCTUBRE DEL 2024

PORFIRIO ROBERTO NÁJERA MEDINA
DIRECTOR DEL INSTITUTO

EDUCACIÓN
INSTITUTO TECNOLÓGICO DE ZACATEPEC
DIRECCIÓN

SC-037/24

Anexo A.2: Ponencia titulada: “Futuro digital: Explorando MEDIAPIPE y Yolo”.

Anexo B: Publicación de Artículo Científico.

Computación y Sistemas

AN INTERNATIONAL JOURNAL OF COMPUTING SCIENCE AND APPLICATIONS

ISSN 1405-5546

Mexico City, 9 December, 2024.

Apartado Postal 75-546
C.P. 07738 México, D.F.
Tel. (+52)-55-57-29-60-00
Ext. 56518, 56643
Fax Ext 56607
computacion-y-sistemas@cic.ipn.mx

http://cys.cic.ipn.mx

ACCEPTANCE OF ARTICLE

This is to confirm that

Carlos Javier Martínez Joffroy

the author of the paper entitled:

"PromptIQ: A New Metric for Evaluating Prompt Quality"

during the double blind reviewing process for the journal Computacion y Sistemas, on November 27, 2024.

Sincerely.

Ph. Grigori Sidorov
Editor in Chief

Anexo B.1: Artículo Publicado en la Revista: "Computación y Sistemas".