Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Extracción de información de datos de texto libre

Tesis de Licenciatura en Ciencias de Datos

Ignacio Rodríguez Sañudo

Director: Germán Rosati
Codirector: Juan Kamienkowski
Buenos Aires, 2025

RESUMEN

La extracción de información a partir de datos de texto libre es un proceso fundamental para transformar grandes volúmenes de información no estructurada en formatos organizados y analizables. Esta tarea adquiere especial relevancia en dominios específicos, como el análisis de denuncias y testimonios, donde el vocabulario, la semántica y los formatos de redacción presentan particularidades. Sin embargo, la aplicación de técnicas de extracción de información en estos contextos se enfrenta a desafíos como la escasez de datos anotados, limitaciones en los recursos computacionales y la necesidad de proteger la privacidad de la información sensible.

Este trabajo se centra en evaluar la factibilidad de emplear modelos de lenguaje autorregresivos para la extracción de información en textos en español provenientes del dominio de denuncias y testimonios. Se investiga el rendimiento de tres variantes de la familia Gemma 3 (con 4, 12 y 27 mil millones de parámetros), exploradas como una alternativa viable para su ejecución local, frente a restricciones de privacidad y recursos. Además, se utiliza el modelo Gemini 2.5 Flash, que permite contrastar el desempeño de un modelo comercial de acceso remoto frente a soluciones que podrían implementarse de forma local.

Para este estudio, se definieron dos tareas principales de extracción: en primer lugar, la extracción de entidades correspondientes a las categorías PERSONA, LUGAR, ORGANIZACIÓN y FECHA. Esta tarea se abordó utilizando enfoques de zero-shot y one-shot learning, con prompts diseñados tanto en español como en inglés, con el objetivo de evaluar el impacto del idioma del prompt en el rendimiento de los modelos. En segundo lugar, se llevó a cabo la detección de eventos de tipo CAPTURA y ASESINATO, empleando estrategias de one-shot learning con prompts en español. La evaluación de ambas tareas se realizó sobre un corpus de 64 documentos del Proyecto Angelus de México, utilizando métricas de precisión, recall y F1-score.

Los resultados demuestran una mejora consistente en el rendimiento a medida que aumenta el tamaño del modelo, siendo la variante Gemma 3 de 27 mil millones de parámetros la que obtiene los mejores resultados dentro de las opciones de tamaño moderado. En la tarea de extracción de entidades, este modelo alcanza un rendimiento comparable, aunque ligeramente inferior al de Gemini 2.5 Flash. En la extracción de eventos, Gemini 2.5 Flash obtiene, con diferencia, mejores resultados que todas las variantes de Gemma 3.

Se concluye que la utilización de modelos de lenguaje autorregresivos, como la variante Gemma 3 de 27 mil millones de parámetros, resulta factible y ofrece un rendimiento prometedor para la extracción de entidades en textos de denuncias y testimonios, especialmente en contextos con restricciones de recursos y privacidad. Sin embargo, para tareas de mayor complejidad, como la extracción de eventos, puede ser conveniente emplear modelos de mayor escala, como Gemini 2.5 Flash, que demuestra una clara superioridad en desempeño.

Palabras claves: extracción de información, modelos de lenguaje, procesamiento del lenguaje natural, extracción de entidades nombradas, extracción de eventos, instruction tuning, in-context learning, Gemma 3, Gemini 2.5, dominio específico, privacidad de datos.

ABSTRACT

Information extraction from free text data is a fundamental process for transforming large volumes of unstructured information into organized and analyzable formats. This task is particularly relevant in specific domains, such as the analysis of complaints and testimonies, where vocabulary, semantics, and writing styles present unique characteristics. However, applying information extraction techniques in these contexts faces challenges such as the scarcity of annotated data, limitations in computational resources, and the need to protect sensitive information privacy.

This work focuses on evaluating the feasibility of using autoregressive language models for information extraction in Spanish texts from the domain of complaints and testimonies. We investigate the performance of three variants of the Gemma 3 family (with 4, 12, and 27 billion parameters), explored as a viable alternative for local execution, given privacy and resource constraints. Additionally, the Gemini 2.5 Flash model is used to contrast the performance of a remote-access commercial model against solutions that could be implemented locally.

For this study, two main extraction tasks were defined: first, the extraction of entities corresponding to the categories PERSON, PLACE, ORGANIZATION, and DATE. This task was addressed using zero-shot and one-shot learning approaches, with prompts designed in both Spanish and English, to evaluate the impact of the prompt language on model performance. The second task involved event detection for CAPTURE and MURDER types, performed using one-shot learning strategies with Spanish prompts. Both tasks were evaluated on a corpus of 64 documents from the Proyecto Angelus in Mexico, using precision, recall, and F1-score metrics.

The results show a consistent improvement in performance as the model size increases, with the Gemma 3 variant of 27 billion parameters achieving the best results among the moderately sized options. In the entity extraction task, this model achieves comparable, though slightly inferior, performance to Gemini 2.5 Flash. In event extraction, Gemini 2.5 Flash significantly outperforms all Gemma 3 variants.

It is concluded that the use of autoregressive language models, such as the Gemma 3 variant with 27 billion parameters, is feasible and offers promising performance for entity extraction in texts from complaints and testimonies, especially in contexts with resource and privacy constraints. However, for more complex tasks, such as event extraction, it may be more convenient to use larger-scale models, such as Gemini 2.5 Flash, which demonstrates clear superiority in performance.

Keywords: information extraction, language models, natural language processing, named entity recognition, event extraction, instruction tuning, in-context learning, Gemma 3, Gemini 2.5, specific domain, data privacy.

Índice general

1. INTRODUCCIÓN

El proceso de extracción de información consiste en convertir, de manera automática, la información contenida en textos a una representación estructurada [1]. Esta transformación acelera procesos que serían muy lentos de realizar manualmente y favorece la integración del conocimiento de distintas fuentes. Además, permite realizar análisis como minería de datos o descubrimiento de patrones que van más allá de las técnicas de procesamiento del lenguaje natural.

La información extraída puede estructurarse de diferentes formas: pueden extraerse entidades (como nombres de personas, organizaciones, ubicaciones o fechas), las relaciones que se establecen entre ellas (por ejemplo, afiliación o vínculos familiares) y los eventos que involucran a dichas entidades. También pueden identificarse atributos específicos asociados a entidades y eventos.

Esta representación estructurada de la información extraída suele darse en distintos formatos según el tipo de datos y los objetivos del análisis. Los formatos más utilizados son las tablas, que pueden almacenar las entidades, los eventos y sus atributos de manera organizada, y los grafos, que permiten modelar las relaciones entre entidades.

En los últimos años, los modelos de lenguaje generativos de gran escala (LLMs, por sus siglas en inglés) han demostrado capacidades notables en la comprensión y generación de texto. Estos avances han impulsado una creciente investigación sobre su aplicación en tareas de extracción de información, enfocándose en aprovechar su potencial para generar salidas estructuradas que contengan la información relevante para cada tarea específica [2].

Si bien un modelo entrenado con grandes cantidades de texto puede ofrecer buenos resultados generales de comprensión y generación de texto, su rendimiento puede disminuir significativamente al aplicarse a un dominio específico. Cada dominio posee un vocabulario propio, una semántica particular y distintos formatos de redacción. Por esto, resulta fundamental la adaptabilidad de las técnicas a las particularidades del campo, tanto en términos de rendimiento como en el formato de salida.

Este trabajo se centra en un caso específico de aplicación: la extracción de información de fuentes textuales en castellano vinculadas a denuncias y testimonios. En este dominio existen tres condicionantes fundamentales: escasez de datos anotados, recursos computacionales limitados y cuestiones vinculadas a la confidencialidad de la información. Estas restricciones condicionan el enfoque hacia el uso de modelos de tamaño moderado, que puedan ejecutarse localmente sin depender de proveedores externos, y de técnicas que no requieran reentrenamiento.

El objetivo principal de este trabajo es evaluar la factibilidad de emplear modelos de lenguaje autorregresivos de gran escala para la extracción de información, poniendo foco en la detección de entidades y eventos definidos según criterios particulares del dominio de estudio. A su vez, el trabajo funciona como prototipo para la tarea de extracción de información en el marco de un proyecto impulsado por Abuelas de Plaza de Mayo, la Facultad de Ciencias Exactas y Naturales de la Universidad de Buenos Aires y la empresa Quantit, que busca aplicar técnicas de inteligencia artificial a la búsqueda y restitución de nietos y nietas apropiados durante la última dictadura militar [3].

La presente tesis se organiza de la siguiente manera.

El Capítulo 2, titulado "Fundamentos", establece el marco teórico necesario, abordando conceptos clave como la extracción de información, los grafos de conocimiento, las ontologías y los modelos de lenguaje autorregresivos junto con la arquitectura Transformer, incluyendo paradigmas como el preentrenamiento y ajuste, el aprendizaje en contexto y el ajuste de instrucciones.

El Capítulo 3, "Dominio de aplicación y trabajos relevantes", describe el contexto específico de las denuncias y documentos similares en español y revisa investigaciones previas tanto en la extracción de información en este dominio como en el uso de modelos autorregresivos para tareas de extracción.

El Capítulo 4, "Metodología", detalla el enfoque experimental, incluyendo la descripción del corpus de datos utilizado, los modelos seleccionados, la implementación de las estrategias de prompting (zero-shot y one-shot) y el proceso de evaluación manual y las métricas empleadas.

El Capítulo 5, "Resultados", presenta y analiza los hallazgos obtenidos.

El Capítulo 6, "Conclusiones y trabajo futuro", discute las implicaciones de los resultados e identifica posibles desarrollos y mejoras para profundizar en la investigación.

Finalmente, el Anexo, recopila los prompts utilizados en la experimentación para las distintas configuraciones. La tesis concluye con la bibliografía consultada.

2. FUNDAMENTOS

A continuación, se presenta una revisión sucinta de algunos conceptos y técnicas fundamentales para abordar el problema de la tesis.

2.1. Extracción de entidades nombradas

La extracción de entidades nombradas (NER, por sus siglas en inglés) consiste en identificar y clasificar automáticamente menciones de entidades específicas dentro de un texto. Los tipos de entidades incluyen nombres de personas, organizaciones, ubicaciones y, en muchos casos, la tarea es extendida a fechas, cantidades monetarias, porcentajes, y otras categorías predefinidas según el dominio de aplicación. El proceso involucra dos problemas principales: la detección de los límites de la entidad (identificar dónde comienza y termina) y la clasificación categórica (asignar la etiqueta correspondiente). Esta tarea presenta desafíos significativos debido a la ambigüedad del lenguaje natural, la variabilidad en las formas de mencionar una misma entidad, y la dependencia del contexto para la correcta interpretación. Varias de estas ambigüedades pueden resolverse cuando se dispone de información contextual suficiente.

2.2. Extracción de Relaciones

La extracción de relaciones se enfoca en detectar y caracterizar las relaciones semánticas que existen entre las entidades mencionadas en un texto. Una relación puede ser de parentesco, espacial o de jerarquía, entre otras. Pueden presentarse de forma explícita o implícita. La complejidad de esta tarea radica en la diversidad de formas para expresar una misma relación semántica y en la necesidad de resolver ambigüedades cuando múltiples interpretaciones son posibles.

2.3. Extracción de Eventos

La extracción de eventos busca reconocer, dentro de un texto, los acontecimientos en los que participan entidades específicas, así como los distintos atributos asociados a cada evento, como la ubicación geográfica y el momento temporal en que ocurre. Además, esta tarea puede extenderse a determinar los roles que asumen las entidades involucradas. Es una tarea más compleja, ya que no solo identifica los eventos y las entidades involucradas, sino que además relaciona a los participantes entre sí, asignándoles roles específicos dentro del evento y situándolos en un tiempo y espacio determinados.

2.4. Grafos de conocimiento

Un grafo de conocimientos es una representación estructurada del conocimiento en forma de red dirigida, donde los nodos representan entidades o eventos del mundo real y las aristas codifican relaciones semánticas entre estas entidades. Los grafos de conocimiento facilitan operaciones avanzadas como búsqueda semántica, razonamiento por inferencia, detección de patrones y descubrimiento de conocimiento implícito. En este contexto, los

grafos de conocimientos constituyen uno de los posibles formatos de representación, donde la información extraída de textos se organiza en una estructura que preserva las relaciones semánticas entre las entidades y permite análisis posteriores.

2.5. Ontologías

Mientras que los grafos de conocimientos proporcionan la estructura de red para representar información, las ontologías definen el marco conceptual que determina qué tipos de entidades, relaciones y eventos pueden existir en un dominio específico, así como las restricciones y propiedades que los caracterizan.

2.6. Extracción abierta vs. cerrada

En la extracción cerrada de información se extraen entidades, relaciones y eventos en base a una ontología predefinida guiada por los objetivos específicos del análisis que se desee realizar. Este enfoque permite extraer información estructurada alineada con un marco conceptual formal, garantizando coherencia semántica y facilitando su integración con sistemas de conocimiento existentes. Además, la información extraída puede ser consultada directamente mediante consultas formales sobre bases de datos o grafos de conocimiento, facilitando la recuperación precisa y sistemática de datos específicos según criterios definidos.

En contraste, el objetivo de la extracción abierta de información es identificar patrones de "sujeto-predicado-objeto" en el texto de manera automática, sin imponer categorías semánticas específicas. Esto permite descubrir información no anticipada, aunque a costa de una menor coherencia semántica y mayores desafíos para su integración en formatos estructurados.

2.7. Modelos de lenguaje autorregresivos

Un modelo de lenguaje autorregresivo asigna una probabilidad a cada token prediciendo su valor en función de los tokens anteriores en la secuencia. El término token hace referencia a una unidad básica de texto procesada por el modelo, que puede corresponder a una palabra completa, una subpalabra o un carácter. Cada token se representa internamente mediante un embedding, que es un vector numérico denso que captura información semántica y sintáctica sobre ese token. La secuencia completa de tokens se organiza en una matriz $X \in R^{n \times d_{modelo}}$, donde n es la cantidad de tokens en la secuencia y d_{modelo} es la dimensión de los embeddings. Por su parte, el contexto se refiere al conjunto de tokens que preceden (y en algunos modelos también a los suceden) a una palabra específica y que influyen en su significado, interpretación o predicción.

La probabilidad de una secuencia completa de tokens se obtiene multiplicando la probabilidad de cada token dado el contexto de los tokens anteriores (Ecuación 2.1).

$$P(w_1, w_2, \ldots, w_n) = \prod_{t=1}^{n} P(w_t \mid w_1, w_2, \ldots, w_{t-1}) \tag{2.1}$$

2.8. De las redes neuronales recurrentes a los Transformers

2.8.1. Redes neuronales recurrentes

Las redes neuronales recurrentes [4] fueron el estándar durante muchos años en tareas de modelado del lenguaje natural debido a su capacidad para procesar secuencias de longitud variable. La arquitectura clásica modela un vector $h^{(t)}$ (estado oculto) recurrentemente en función del estado oculto anterior $h^{(t-1)}$ y el token actual $x^{(t)}$ (Ecuaciones 2.2 y 2.3). Luego, se utiliza únicamente este vector para predecir el siguiente token (Ecuaciones 2.4 y 2.5). De esta manera, el estado oculto debe codificar toda la información de los tokens anteriores y las matrices que multiplican a los vectores de cada paso son las mismas.

$$a^{(t)} = b + Wh^{(t-1)} + Ux^{(t)} \tag{2.2}$$

$$h^{(t)} = \tanh(a^{(t)}) \tag{2.3}$$

$$o^{(t)} = c + Vh^{(t)} \tag{2.4}$$

$$\hat{y}^{(t)} = softmax(o^{(t)}) \tag{2.5}$$

Esto genera dos dificultades significativas: la primera es que el estado oculto y los parámetros que determinan sus valores deben proveer simultáneamente información para el siguiente y para todos los subsiguientes tokens. La segunda es el desvanecimiento del gradiente al retropropagar el error a través del tiempo. Durante la fase de retropropagación, el estado oculto influye en la función pérdida del siguiente paso temporal. Por lo tanto las capas ocultas son sometidas a múltiples multiplicaciones sucesivas, dependiendo de la longitud de la secuencia. Esto con frecuencia provoca que los gradientes se reduzcan progresivamente hasta llegar a valores cercanos a cero. [1]

Para atenuar estas limitaciones, se desarrollaron las Long Short-Term Memory (LSTM) [5], que introducen mecanismos de compuertas para controlar el flujo de información al remover o incorporar información del contexto en cada paso. No obstante, aunque las LSTM representan un avance significativo, ambos problemas persisten en menor medida y limitan la captura de dependencias prolongadas.

En base a estas arquitecturas, se desarrollaron los modelos encoder-decoder, capaces de generar secuencias de salida de longitud arbitraria a partir de una secuencia de entrada. La idea central de este enfoque es utilizar un encoder que procesa la secuencia de entrada y genera una representación contextualizada de la misma. Esta representación se transmite a un decoder, encargado de generar la secuencia de salida según la tarea específica [1].

2.8.2. Mecanismo de atención

Más adelante, se introdujeron los modelos basados en atención [6] que permiten que el decoder se enfoque directamente en partes relevantes de la secuencia de entrada al calcular un vector contextual en base a la similaridad entre todos los estados ocultos del encoder y el estado oculto anterior del decoder (Ecuaciones 2.6 y 2.7) para finalmente definir el estado oculto del decoder en función de la predicción anterior, el estado oculto anterior y el vector contextual, como se muestra en la Ecuación 2.8. Como resultado, se logra superar

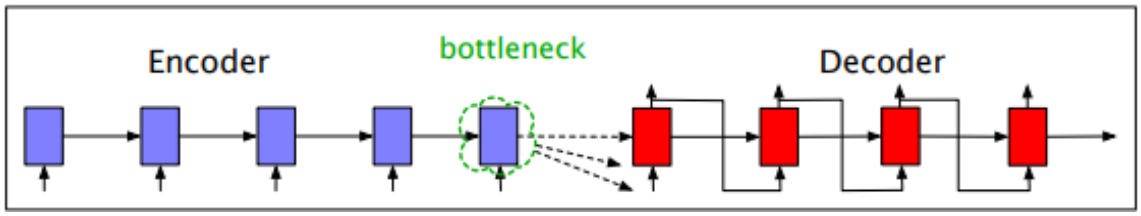

Fig. 2.1: Cuello de botella en redes neuronales recurrentes. Toda la información de la entrada que recibe el decoder viene del último estado oculto del encoder. [1]

las restricciones del flujo de la información contextual en las redes neuronales recurrentes tradicionales.

$$\alpha_{ij} = softmax\left(\mathbf{h}_{i-1}^{d}\mathbf{W}_s\mathbf{h}_j^e\right) \tag{2.6}$$

$$c_i = \sum_j \alpha_{ij}\mathbf{h}_j^e \tag{2.7}$$

$$\mathbf{h}_i^d = g(\hat{y}_{i-1}, \mathbf{h}_{i-1}^d, \mathbf{c}_i) \tag{2.8}$$

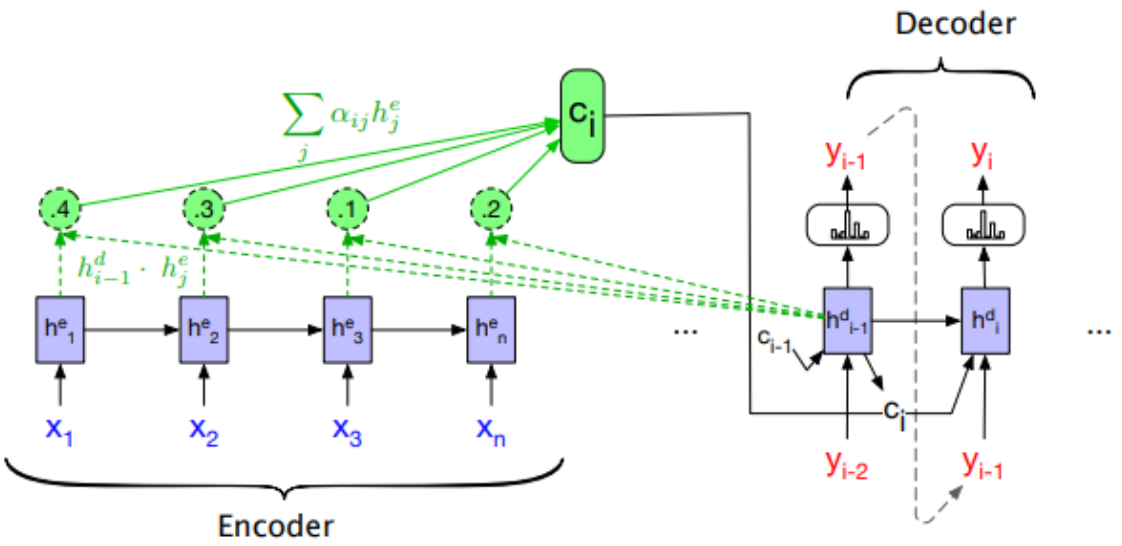

Fig. 2.2: Diagrama de una red neuronal recurrente con el mecanismo de atención. [1]

2.8.3. Transformers

Desde su introducción, los modelos basados en arquitecturas de Transformers [7], y en particular los modelos autorregresivos, se han consolidado como estándar en el procesamiento automático de lenguaje natural debido a su notable capacidad para modelar secuencias y su alto rendimiento en una amplia variedad de tareas.

Esta arquitectura incorpora el mecanismo de autoatención (self-attention) como la única forma de capturar dependencias globales entre tokens dentro de una secuencia, abandonando la recurrencia y permitiendo que el modelo asigne dinámicamente importancia a

diferentes posiciones sin importar su ubicación y capture relaciones contextuales de largo alcance.

A diferencia de las redes neuronales recurrentes, que procesan las palabras de forma secuencial, los Transformers procesan las palabras en paralelo, lo que les brinda mayor eficiencia computacional y una mejor capacidad para captar relaciones de largo alcance en el texto.

El mecanismo de Scaled Dot-Product Attention produce una representación para cada token como una combinación lineal ponderada de las representaciones de todos los tokens de la secuencia (V), donde los pesos se determinan mediante un softmax aplicado al producto punto escalado entre (Q) y las claves (K) (Ecuación 2.9). Estos pesos son mayores cuando las representaciones Q y K de dos tokens son más similares, ya que su producto punto resulta en un valor mayor, lo que indica una mayor relevancia contextual entre esas posiciones. A las matrices X se les aplican transformaciones lineales para obtener Q, K y V (Ecuación 2.10).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.9}$$

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{2.10}$$

Para captar diferentes tipos de dependencias entre los tokens y enriquecer la comprensión contextual del modelo, la arquitectura de Transformer puede concatenar en un sólo vector para cada token múltiples mecanismos de atención con diferentes matrices de proyección para una misma entrada X. Luego, se aplica una proyección lineal sobre este resultado. Esto se conoce como Multi-Head Attention (Fig. 2.3).

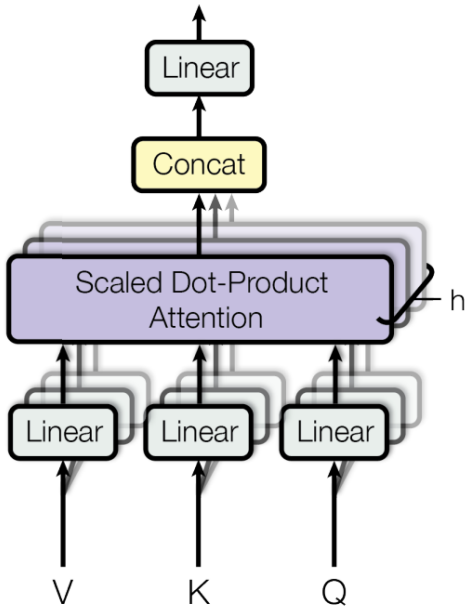

Fig. 2.3: Multi-Head Attention. [7]

La arquitectura Transformer se diseñó originalmente para tareas de secuencia a secuencia y está formada por dos componentes principales: un Encoder y un Decoder, cada

uno compuesto por varias capas apiladas (Fig. 2.4).

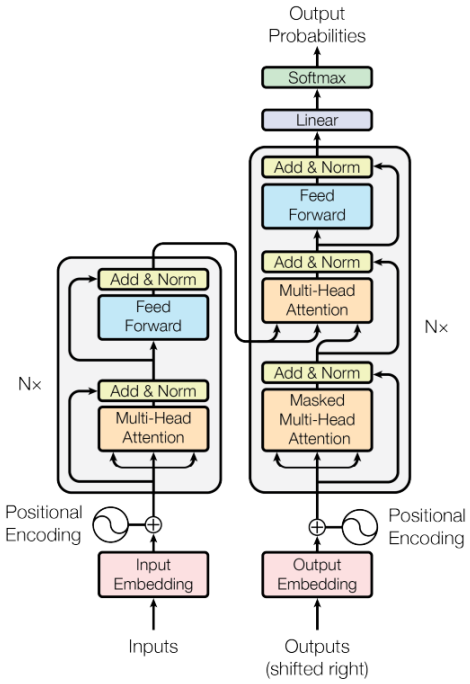

Fig. 2.4: Arquitectura del Transformer. [7]

El Encoder es responsable de procesar la secuencia de entrada para generar una representación contextual de la misma. Cada capa contiene dos submódulos: un mecanismo de multi-head self-attention y una red feed-forward que se aplica de forma independiente a cada posición.

El Decoder se encarga de generar la secuencia de salida de manera autorregresiva. Además de los dos submódulos que tiene el encoder, cada capa del decoder incluye un tercer submódulo de multi-head attention, que toma como Q las salidas de la capa previa del decoder, y como K y V las salidas finales del encoder. El mecanismo de self-attention del decoder incluye un enmascaramiento de la atención que impide que cada posición acceda a información de posiciones futuras, asegurando así la generación secuencial.

Ambos módulos incorporan conexiones residuales y normalización de capas para facilitar el entrenamiento.

Dado que el mecanismo de atención es invariante al orden de los tokens, se suma una codificación posicional a las representaciones de entrada para proporcionar información sobre la posición relativa o absoluta de cada token dentro de la secuencia (Ecuaciones 2.11 y 2.12).

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{modelo}}}}\right) \tag{2.11}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{modelo}}}}\right) \tag{2.12}$$

La capacidad de los Transformers de modelar relaciones entre palabras no contiguas, junto con una mayor eficiencia computacional gracias a la paralelización y a la reducción

de operaciones secuenciales, ha sido fundamental para el éxito de LLMs. Según la tarea, la arquitectura Transformer puede adaptarse utilizando únicamente el encoder, como en BERT [8], o solo el decoder, como en GPT [9]. En este último caso, se elimina la subcapa de encoder-decoder attention, ya que el modelo no necesita considerar el embedding del encoder, sino únicamente generar texto de manera autorregresiva. Estas variantes especializadas son la base de la mayoría de los modelos de lenguaje actuales.

Además de su diseño, el éxito de los Transformers se debe también a los enfoques de entrenamiento empleados para aprovechar su flexibilidad y capacidad de modelado contextual. El paradigma *pretraining-finetuning* aplicado a texto [9] ha sido fundamental para el avance de los modelos de lenguaje. En este enfoque, el modelo se entrena inicialmente en grandes cantidades de texto sin etiquetas para aprender representaciones generales del lenguaje (*pretraining*), y posteriormente se ajusta con datos anotados específicos para tareas concretas (*finetuning*). Conjuntamente, han surgido nuevos paradigmas como el *in-context learning* [10] y el *instruction tuning* [11], los cuales se manifiestan principalmente en modelos basados en arquitecturas Transformer, dada su capacidad para manejar de forma flexible amplios contextos y adaptarse dinámicamente a instrucciones específicas.

2.9. In-Context Learning

El in-context learning es una capacidad emergente de los modelos de lenguaje basados en Transformers, en la que el modelo puede resolver tareas para las que no fue entrenado sin modificar sus parámetros. Consiste en utilizar la entrada textual de un modelo de lenguaje preentrenado como una forma de especificación de tarea. En este paradigma, el modelo se condiciona mediante una instrucción en lenguaje natural y/o mediante algunas demostraciones o ejemplos de la tarea, y se espera que complete nuevas instancias simplemente prediciendo la continuación más probable. Este procedimiento no requiere modificar los parámetros del modelo, ya que el aprendizaje ocurre a partir del contexto incluido en la propia secuencia de entrada. Generalmente se distinguen tres configuraciones: **few-Shot**, donde el modelo recibe algunas demostraciones de la tarea en la entrada como ejemplo; **one-Shot**, en el que se proporciona una única demostración junto a una instrucción en lenguaje natural; y **zero-Shot**, donde solo se incluye la instrucción sin ningún ejemplo adicional.

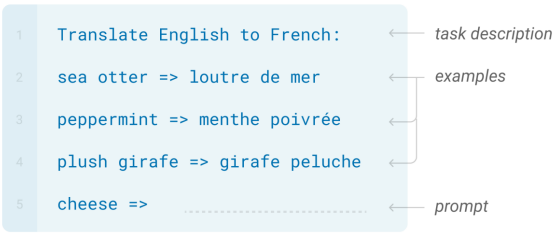

Fig. 2.5: Entrada textual en la que se proporcionan algunos ejemplos (few-shot learning). [10]

2.10. Instruction Tuning

Instruction tuning es una técnica de ajuste supervisado aplicada a modelos de lenguaje preentrenados, que consiste en entrenarlos con un gran conjunto de pares del tipo:

$$\langle instrucción, respuesta\ esperada \rangle$$

El objetivo del instruction tuning es mejorar la capacidad del modelo para seguir instrucciones formuladas en lenguaje natural de manera generalizada. Este enfoque se basa en el paradigma clásico de pretraining-finetuning, ya que implica modificar los parámetros del modelo. Reduce significativamente la necesidad de proporcionar ejemplos en el momento de la inferencia, dado que el modelo ha aprendido a interpretar y resolver tareas diversas a partir de instrucciones explícitas.

Un tipo de respuesta esperada en *instruction tuning* puede incluir no solo la solución final, sino también los pasos intermedios de razonamiento en lenguaje natural, lo que se conoce como *Chain-of-Thought (CoT)*. Este enfoque mejora significativamente la capacidad de los modelos para resolver tareas de razonamiento complejo [11]. La misma lógica se puede aplicar a los ejemplos proporcionados durante la inferencia mediante *in-context learning*, obteniendo mejoras similares sin necesidad de ajustar los parámetros del modelo [12].

3. DOMINIO DE APLICACIÓN Y TRABAJOS RELEVANTES

3.1. Dominio de aplicación

En el marco de un proceso investigativo orientado a la reconstrucción de hechos, resulta fundamental apoyarse en la información contenida en distintos tipos de textos. Estos documentos, que pueden provenir de declaraciones, entrevistas, reportes, denuncias formales y otras fuentes documentales, suelen presentarse en formatos diversos y emplear un lenguaje particular dado por el contexto en que fueron elaborados.

La extracción automática de información en este tipo de escenarios resulta crucial para facilitar la organización, sistematización y análisis de grandes volúmenes de datos textuales, permitiendo acelerar tareas que, realizadas manualmente, serían extensas y demandantes. Además, posibilita la integración de información dispersa y fragmentada, contribuyendo a construir una visión más completa y precisa de los hechos investigados. Este tipo de técnicas no solo permite recuperar información explícita, sino también identificar entidades típicamente relevantes en procesos investigativos, como personas, lugares, fechas, organizaciones y objetos, y sugerir relaciones entre ellas. Asimismo, facilita la detección de patrones, conexiones entre los hechos y pistas que podrían resultar determinantes para orientar o destrabar una investigación.

La información contenida en este tipo de textos suele ser sensible, ya que puede incluir datos personales, testimonios privados y otros contenidos que requieren un tratamiento confidencial y no deben ser expuestos en entornos públicos o inseguros. Esto constituye un desafío, ya que puede limitar la posibilidad de utilizar los modelos más potentes, que además de requerir un considerable poder de cómputo, suelen operar en entornos externos, lo cual podría comprometer la confidencialidad de la información.

3.2. Trabajos relevantes

3.2.1. Extracción de información en el dominio

Un trabajo fundamental en el ámbito de los modelos de lenguaje aplicados al dominio legal es el que introduce MultiLegalPile, un corpus multilingüe de 689 GB compuesto por textos jurídicos en múltiples idiomas [13]. Este corpus constituye una base de datos crítica para el desarrollo de modelos de lenguaje en el ámbito legal, al proporcionar material especializado para el preentrenamiento de modelos. Sobre este corpus, los autores del mismo preentrenaron diversas arquitecturas, incluyendo versiones de RoBERTa, un modelo basado en la arquitectura BERT, y Longformer, una variante de la arquitectura Transformer diseñada para manejar secuencias extensas mediante mecanismos de atención local y global combinados. Como resultado, lograron establecer un nuevo estado del arte (SOTA) en tareas específicas del dominio jurídico, evidenciando el impacto de disponer de un corpus especializado y arquitecturas adaptadas para abordar los desafíos de longitud y terminología en textos legales.

En el ámbito de la extracción de información sensible aplicada a contextos históricos y sociales, existe el modelo de NER desarrollado por la Comisión de la Verdad de Colombia [14]. Este modelo fue entrenado sobre un corpus extenso de entrevistas transcritas y etiquetadas manualmente, centrado en el conflicto armado colombiano. Su propósito es

automatizar procesos de anonimización documental, lo que permite publicar información sensible sin exponer datos personales, y además optimiza la búsqueda y análisis de datos históricos mediante el etiquetado de entidades como personas, ubicaciones y fechas relevantes.

Con un enfoque similar, AymurAI [15] es una herramienta de inteligencia artificial interactiva creada para asistir a funcionarios judiciales en la gestión de información sobre violencia de género en América Latina. Mediante un modelo de NER, es posible organizar información de manera estructurada y, así, reducir la dificultad de acceso y uso de los datos existentes, además de facilitar la apertura de datos previamente anonimizados sobre violencia de género.

Otro ejemplo de modelo especializado de clasificación de texto en español es MEL (Modelo Español Legal), un modelo de lenguaje encoder-only basado en XLM-RoBERTa-large, una versión multilingüe optimizada para tareas en múltiples idiomas y ajustada en este caso exclusivamente para castellano jurídico. MEL fue preentrenado sobre una colección de más de 40 GB de textos jurídicos en español [16]. Su diseño busca abordar los desafíos lingüísticos asociados a la terminología técnica y la estructura argumentativa de documentos legales. Los resultados de su evaluación demuestran que MEL supera ligeramente el rendimiento de modelos multilingües preexistentes en tareas de procesamiento del lenguaje natural del ámbito legal en español.

Por otra parte, aunque no se apoya en métodos de extracción automática, la relevancia de la sistematización y organización de información sensible en contextos históricos y sociales se manifiesta en iniciativas como el Proyecto Angelus [17]. Este proyecto, enfocado en la desaparición forzada de personas en México durante la "Guerra sucia", se centra en la definición y formalización de una ontología computacional diseñada para representar y estructurar las relaciones entre personas, eventos, lugares y roles. Con el objetivo principal de crear una base de datos estructurada a partir de fuentes primarias y secundarias, el Proyecto Angelus facilita la búsqueda de verdad y la memoria histórica a partir de un corpus de datos extenso que incluye documentos de archivos históricos, testimonios y otras fuentes relevantes.

3.2.2. Extracción de información con modelos autorregresivos

La robustez de los LLMs frente a datos ruidosos o específicos del dominio es crucial. En NER4all [18] abordan la extracción de entidades nombradas en textos históricos, destacando que el contexto y el modelado de las instrucciones son fundamentales para lograr un alto rendimiento en dominios con variabilidad lingüística y escasez de datos.

En la tarea de extracción de relaciones, un ejemplo relevante es el trabajo de Zhang y Soh (2024) [19], quienes proponen un marco de tres fases (Extract, Define, Canonicalize) para la construcción de grafos de conocimiento. Cada una de estas fases se implementa mediante un prompt (entrada textual del modelo) con instrucciones específicas al modelo, demostrando cómo los modelos generativos pueden generar esquemas de forma automática y extraer triplas, incluso cuando se trabaja con esquemas u ontologías complejos.

En esta misma línea, Edge et al. (2024) presentan GraphRAG [20], una herramienta que construye un grafo de conocimiento sin una ontología predefinida, aplica un agrupamiento jerárquico para identificar comunidades dentro del grafo, genera resúmenes para cada una de ellas y utiliza esta estructura para responder preguntas sobre el corpus. La construcción del grafo, la elaboración de resúmenes y la generación de respuestas se basan en instrucciones y ejemplos proporcionados a un LLM.

Un ejemplo reciente de aplicación de LLMs a la extracción de eventos es el estudio de Lu et al. (2025) [21], quienes investigan el uso de estos modelos para identificar eventos y sus características. En el trabajo se utilizan métodos de evaluación semántica que van más allá de la coincidencia exacta al emplear otros LLMs avanzados como agentes evaluadores para comparar las extracciones producidas por el modelo con las anotaciones y determinar su corrección en función del contexto.

4. METODOLOGÍA

4.1. Datos

4.1.1. Descripción

Para los fines de esta tesis, se utilizó un subconjunto de 64 documentos digitalizados provenientes de un conjunto de datos público del Proyecto Angelus [22]. Este conjunto forma parte de los documentos base empleados en la construcción de los grafos de conocimiento desarrollados por el proyecto.

Los documentos corresponden a copias maestras de expedientes resguardados en el Archivo General de la Nación (México), generados en el contexto de operaciones de la llamada contrainsurgencia. Incluyen textos en español de diversas fuentes, como denuncias, testimonios y artículos periodísticos.

Se seleccionaron estos 64 documentos específicamente por su extensión y variabilidad, con el propósito de contar con un corpus representativo y diverso que permitiera evaluar las estrategias de análisis propuestas.

4.1.2. Preprocesamiento

Los documentos fueron procesados mediante reconocimiento óptico de caracteres (OCR) utilizando el modelo Gemini 2.5 Pro [23], con el propósito de convertir los archivos digitalizados en texto editable para su posterior análisis.

Este procedimiento, como es habitual en procesos de OCR, puede introducir errores de reconocimiento, especialmente en documentos con baja calidad de imagen o deterioro físico. Entre los errores más frecuentes se detectaron saltos de línea que interrumpían palabras o fragmentaban párrafos de manera incorrecta en los textos generados, afectando la coherencia del texto resultante. Algunos de estos errores fueron identificados y corregidos manualmente durante una fase de revisión del corpus.

4.2. Modelos

4.2.1. Elección de los modelos

Para la implementación de este trabajo, se seleccionaron cuatro modelos multimodales: tres de la familia Gemma 3 [24], en sus variantes Gemma-3-4b-it, Gemma-3-12b-it y Gemma-3-27b-it, con entre 4 y 27 mil millones de parámetros, y el modelo comercial Gemini 2.5 Flash [25].

La elección de los modelos de la familia Gemma 3 responde directamente a los condicionantes de este estudio. Su tamaño moderado permite una ejecución eficiente en entornos con recursos computacionales limitados, lo que facilita su ejecución local y, por ende, garantiza la confidencialidad de la información. Además, su capacidad para interpretar y seguir instrucciones explícitas resulta fundamental para afrontar la escasez de datos anotados en el dominio abordado.

Adicionalmente, se incluyó Gemini 2.5 Flash por su desempeño en tareas de razonamiento complejo [25].

La evaluación de estas cuatro variantes permite analizar cómo la escala del modelo incide en la calidad y precisión en la tarea de extracción de entidades en un dominio específico.

4.2.2. Gemma 3

En el reporte técnico de los modelos [24], se da a conocer la arquitectura, la cantidad parámetros y ciertos detalles sobre su implementación. Sin embargo, no toda la información relacionada con su entrenamiento y configuración ha sido publicada.

Los modelos Gemma 3 basan su arquitectura en los Transformers, utilizando únicamente el decoder. Aunque el foco principal es el texto, permiten procesar imágenes representándolas como una secuencia de tokens. Esta codificación inicial no forma parte del entrenamiento, sino que los modelos se entrenan para aprender a utilizar la representación inicial junto con el texto.

Además, estos modelos presentan diferencias importantes respecto a los Transformers originales, incorporando modificaciones como Grouped-Query Attention (GQA), donde varios grupos de consultas (Q) comparten las mismas claves (K) y valores (V) dentro del mecanismo de multi-head attention (Ecuación 2.9). También implementan QK-norm, una técnica que normaliza el producto QK mediante un escalar entrenable, lo que contribuye a estabilizar y optimizar el entrenamiento. Una característica clave que mejora notablemente la eficiencia computacional es la intercalación de capas de atención local y global, en la que se alternan capas con contextos de 128 mil tokens y capas con contextos de 1024 tokens para cada token, lo que permite manejar secuencias extensas sin un incremento desproporcionado en el uso de memoria.

Estas versiones fueron entrenadas mediante técnicas de instruction tuning, lo cual las hace adecuadas para este tipo de tareas. El proceso de instruction tuning se realiza en una fase de posentrenamiento mediante una combinación de técnicas: se utiliza destilación de conocimiento, que es el proceso de transferir el conocimiento de un modelo grande y complejo a uno más pequeño y simple, a partir de un modelo maestro especializado en seguir instrucciones, junto con técnicas de aprendizaje por refuerzo con retroalimentación humana (RLHF) y modelos de recompensa ponderados. Este enfoque permite optimizar no solo la capacidad del modelo para seguir instrucciones, sino también mejorar habilidades específicas como razonamiento matemático, generación de código, capacidades en múltiples idiomas y minimizar respuestas potencialmente dañinas.

Los modelos Gemma 3 han demostrado un rendimiento competitivo en comparación con modelos de mayor tamaño en una amplia variedad de tareas, como razonamiento, matemáticas y comprensión del lenguaje, evaluadas en benchmarks de referencia como MMLU-Pro, LiveCodeBench, entre otros.

4.2.3. Gemini 2.5 Flash

El reporte técnico de los modelos Gemini 2.5 [25] es menos detallado. Su arquitectura está basada en *sparse Mixture-of-Experts (MoE)*: se entrenan conjuntamente modelos expertos en distintas tareas y un modelo (router) que decide qué expertos son idóneos para una determinada entrada. Como resultado, se activa un subconjunto de expertos según las características de cada token, reduciendo el número de parámetros utilizados en cada paso de inferencia y logrando una mayor eficiencia. La salida de los expertos seleccionados

se combina, generalmente, mediante una suma ponderada, para producir la representación final utilizada en la predicción.

Para el *instruction tuning*, también se utiliza destilación de conocimiento combinada con técnicas de aprendizaje por refuerzo, con énfasis en mejorar la calidad y seguridad de las respuestas. Además, los modelos *Gemini 2.5* incorporan entrenamiento con aprendizaje por refuerzo con el objetivo de que el modelo realice una fase de razonamiento. Los modelos pueden realizar múltiples pasos de inferencia durante una fase de razonamiento antes de generar una respuesta definitiva.

Estos modelos cuentan con un contexto de aproximadamente un millón de tokens y pueden procesar texto, imágenes, audio y videos.

Gemini 2.5 Flash se consolida como el segundo modelo más potente de la familia Gemini, únicamente por detrás de Gemini 2.5 Pro. Según los resultados reportados, mantiene un rendimiento competitivo en benchmarks de razonamiento, factualidad, multimodalidad y código, posicionándose como una alternativa equilibrada entre desempeño, costo y latencia. Aunque el número exacto de parámetros no ha sido divulgado, se asume que su cantidad total es superior a la de Gemma 3 27B.

4.3. Implementación

La estrategia consistió en definir con precisión las tareas mediante instrucciones claras que especifican las estructuras a extraer, un conjunto de reglas particulares para guiar el proceso de identificación (buscando reflejar los criterios utilizados en la fase de anotación manual) y un formato de salida JSON para garantizar la uniformidad y la compatibilidad con los procesos de análisis posteriores.

4.3.1. Extracción de entidades

Para llevar a cabo la extracción de entidades, se decidió extraer las clases **PERSONA**, **LUGAR**, **ORGANIZACIÓN** y **FECHA** por su gran representación en el corpus y se diseñó un conjunto de prompts orientados a obtener respuestas estructuradas por parte de los modelos de lenguaje.

Se implementaron dos paradigmas de aprendizaje in-context: one-shot learning, incorporando un ejemplo completo y representativo de la tarea, y zero-shot learning, basado únicamente en instrucciones, con el fin de evaluar la capacidad del modelo para resolver la tarea en contextos con y sin ejemplos previos. También se desarrollaron versiones de los prompts en español e inglés con el objetivo de explorar el rendimiento de los modelos en ambos idiomas, considerando que los textos analizados se encuentran en español. Los prompts empleados se encuentran disponibles en el Anexo.

4.3.2. Extracción de eventos

Para la tarea de extracción de eventos, se decidió identificar los tipos: **CAPTURA** (secuestros o detenciones) y **ASESINATO**. Estos dos tipos de eventos también se encuentran definidos en la ontología del proyecto Angelus junto a varios más. La estructura esperada para cada evento incluye los siguientes campos: *descripción*, *tipo*, *fecha*, *lugar*, *víctimas* y *victimarios*. Para ello, se diseñaron dos prompts en español: uno que exige una descripción del evento como parte de la salida y otro que omite este campo. Los prompts contienen definiciones y reglas específicas de cada campo, un ejemplo de salida esperada y

establecen el objetivo de extracción. Además, dentro del prompt, se adjuntan las entidades previamente extraídas para facilitar la generación de información estructurada en base a las reglas definidas tanto para las entidades como para los eventos. Ambos se encuentran en el Anexo.

4.4. Evaluación

Se realizó una fase de anotación manual sobre los documentos seleccionados. Esta tarea consistió en identificar y etiquetar las menciones de **PERSONAS**, **ORGANIZACIONES**, **LUGARES** y **FECHAS** y eventos de tipos **CAPTURA** y **ASESINATO** junto con sus campos. La anotación se llevó a cabo siguiendo criterios definidos para asegurar la consistencia en la clasificación, contemplando alias, siglas y variantes.

4.4.1. Extracción de entidades

El proceso de evaluación de la extracción de entidades se basó en comparar las predicciones de los modelos con las anotaciones manuales mediante un matching exacto. Este método consiste en considerar como correcta una predicción únicamente si coincide de forma idéntica con la anotación esperada. Para minimizar el impacto de diferencias irrelevantes en la comparación, como variaciones en tildes, mayúsculas, saltos de línea, guiones o caracteres especiales, se aplicó una función de normalización a ambos textos antes de realizar la comparación. Esta función reemplaza saltos de línea y secuencias de escape por espacios, elimina tildes y diéresis, convierte todo a mayúsculas, remueve caracteres especiales no alfanuméricos y normaliza los espacios, dejando únicamente letras, números y puntuación básica. Finalmente, se eliminan todos los espacios con el fin de resolver los errores introducidos por el procesamiento con OCR. De este modo, se asegura que las coincidencias se determinen en función del contenido relevante.

Se calcularon las métricas de precisión, recall y F1 para cada clase individualmente, y un promedio de estas métricas para obtener una evaluación global de los modelo (macro average).

4.4.2. Extracción de eventos

En el caso de la extracción de eventos, se calcularon las métricas de precisión, recall y F1 campo a campo (exceptuando la descripción del evento) para cada par formado por un evento extraído y un evento anotado dentro de un mismo documento. Para la comparación entre los textos de cada campo, se consideró una coincidencia si la proporción entre la longitud de la subsecuencia común más larga (no necesariamente de caracteres contiguos) y la suma de las longitudes de ambos textos superaba el umbral de 0,85.

Luego, para cada evento, se calculó un valor promedio de las métricas a partir de los resultados obtenidos en los distintos campos. Con estos valores, se buscó el emparejamiento entre eventos extraídos y anotados que maximizara el F1 promedio. Los eventos sin emparejar fueron considerados falsos positivos o falsos negativos según correspondiera, afectando a la precisión y al recall respectivamente.

Finalmente, se calcularon las mismas métricas para cada campo teniendo en cuenta los emparejamientos en cada documento.

5. RESULTADOS

5.1. Extracción de entidades

5.1.1. Resultados Generales

En esta sección se presentan los resultados obtenidos de la evaluación de los modelos en la tarea de extracción de entidades (**PERSONA**, **LUGAR**, **ORGANIZACIÓN**, **FECHA**) bajo las diferentes configuraciones.

Para ofrecer una visión global del rendimiento, las Tablas 5.1, 5.2, y 5.3 presentan los valores de precision, recall y F1-score obtenidos por cada variante de modelo, considerando las diferentes configuraciones de idioma y tipo de evaluación.

Idioma	**Estrategia**	**4b**	**12b**	**27b**	**2.5 Flash**
Inglés	one-shot	0.681	0.731	0.843	0.803
	zero-shot	0.651	0.715	0.763	0.766
Español	one-shot	0.687	0.739	0.851	0.809
	zero-shot	0.661	0.730	0.773	0.768

Tab. 5.1: Precision Macro Average.

Idioma	**Estrategia**	**4b**	**12b**	**27b**	**2.5 Flash**
Inglés	one-shot	0.627	0.749	0.809	0.859
	zero-shot	0.629	0.723	0.766	0.812
Español	one-shot	0.668	0.742	0.833	0.885
	zero-shot	0.661	0.707	0.782	0.796

Tab. 5.2: Recall Macro Average.

Idioma	**Estrategia**	**4b**	**12b**	**27b**	**2.5 Flash**
Inglés	one-shot	0.652	0.738	0.825	0.829
	zero-shot	0.636	0.718	0.764	0.786
Español	one-shot	0.676	0.738	0.841	0.844
	zero-shot	0.655	0.718	0.777	0.781

Tab. 5.3: F1-score Macro Average.

Al analizar los resultados globales, se observa una tendencia clara de mejora en el rendimiento cuando aumenta el tamaño del modelo.

En cuanto a la estrategia, los modelos obtienen mejores resultados con one-shot. Esta tendencia se mantiene en casi todas las combinaciones de idioma y tamaño de modelo. La única excepción se da en Recall para el modelo de 4b en inglés, donde zero-shot supera levemente a one-shot.

Comparando los resultados obtenidos según el idioma, en general, los modelos muestran una ventaja en español. Esto puede estar relacionado con características del corpus y con las capacidades multilenguaje de los modelos. En la figura 5.1 se pueden ver estas tendencias para el F1-score.

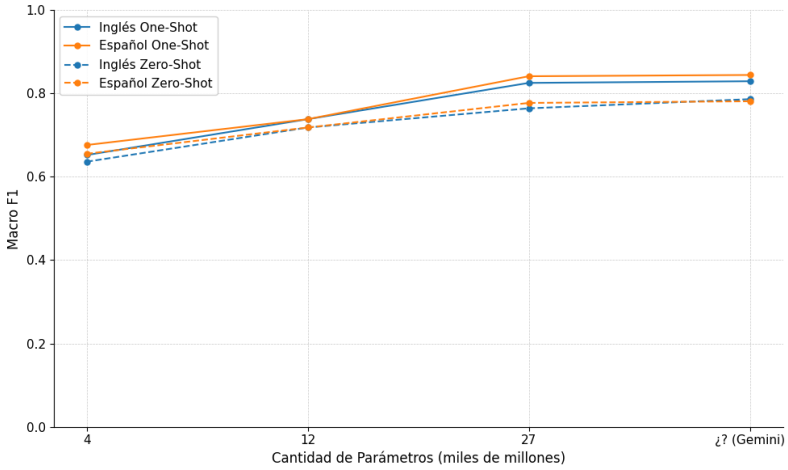

Fig. 5.1: Comparación del F1 Macro Average entre modelos y configuraciones para extracción de entidades.

5.1.2. Resultados detallados por tipo de entidad

Si bien la tendencia general de mejora en el rendimiento con el aumento del tamaño del modelo se mantiene en gran medida al analizar cada tipo de entidad individualmente, el rendimiento comparativo entre los tipos de entidad presenta grandes diferencias.

Los resultados obtenidos para el tipo **PERSONA** (Tabla 5.4) son particularmente robustos, incluso en los modelos de menor tamaño como 4b. Esto podría deberse a que dentro de la clase no se presentan tantas variaciones en la forma de las menciones, lo que facilita su identificación y clasificación consistente por parte de los modelos.

Idioma	**Estrategia**	**Modelo**	**Precision**	**Recall**	**F1-score**
Inglés	one-shot	2.5 Flash	0.920	0.920	0.920
		27b	0.929	0.889	0.909
		12b	0.875	0.852	0.863
		4b	0.814	0.744	0.777
	zero-shot	2.5 Flash	0.943	0.935	0.939
		27b	0.919	0.872	0.895
		12b	0.889	0.864	0.876
		4b	0.833	0.696	0.759
Español	one-shot	2.5 Flash	0.960	0.949	0.954
		27b	0.929	0.889	0.909
		12b	0.884	0.821	0.851
		4b	0.825	0.764	0.794
	zero-shot	2.5 Flash	0.918	0.886	0.902
		27b	0.920	0.881	0.900
		12b	0.882	0.827	0.853
		4b	0.870	0.761	0.812

Tab. 5.4: Resultados para *PERSONA*.

Para las entidades de tipo **LUGAR** (Tabla 5.5) también muestra un rendimiento sólido, posicionándose consistentemente como la segunda entidad mejor reconocida después de **PERSONA**. Aunque los nombres de lugares pueden tener cierta variabilidad (ciudades, países, regiones, direcciones específicas), parecen ser fácilmente delimitables.

Idioma	Estrategia	Modelo	Precision	Recall	F1-score
Inglés	one-shot	Flash 2.5	0.810	0.833	0.821
		27b	0.866	0.870	0.868
		12b	0.779	0.834	0.805
		4b	0.729	0.605	0.661
	zero-shot	Flash 2.5	0.694	0.649	0.671
		27b	0.804	0.779	0.791
		12b	0.748	0.775	0.761
		4b	0.689	0.605	0.644
Español	one-shot	Flash 2.5	0.812	0.880	0.845
		27b	0.871	0.881	0.876
		12b	0.767	0.834	0.799
		4b	0.754	0.668	0.709
	zero-shot	Flash 2.5	0.713	0.693	0.703
		27b	0.807	0.779	0.793
		12b	0.738	0.723	0.731
		4b	0.698	0.585	0.637

Tab. 5.5: Resultados para *LUGAR*.

Por otro lado, para la entidad **FECHA** (Tabla 5.6) se observa un rendimiento significativamente inferior en comparación con las entidades **PERSONA** y **LUGAR**. En el caso del modelo Gemma 3 27b, se aprecia con mayor claridad la influencia de los ejemplos incluidos en los prompts sobre el desempeño obtenido. Estos bajos resultados podrían atribuirse a la especificación de la tarea para este tipo de entidad (una definición muy amplia), y se ve cómo el formato esperado de las FECHAS se completa a partir de los ejemplos proporcionados en el prompt, lo que condiciona la capacidad del modelo para generalizar.

Finalmente, los rendimientos de los modelos para la entidad **ORGANIZACIÓN** se sitúan de forma consistente por debajo de los obtenidos para las otras tres clases en la mayoría de las configuraciones evaluadas. Este comportamiento podría explicarse por la alta variabilidad y complejidad de los nombres de organizaciones, que pueden presentarse en forma de acrónimos, siglas o nombres completos, algunos de los cuales coinciden o se confunden fácilmente con frases completas. Un ejemplo ilustrativo de esto es *Comité Nacional Pro Defensa de Presos Perseguidos, Desaparecidos y Exiliados Políticos, Sección México.*

Idioma	Estrategia	Modelo	Precision	Recall	F1-score
Inglés	one-shot	Flash 2.5	0.745	0.859	0.798
		27b	0.780	0.780	0.780
		12b	0.579	0.684	0.627
		4b	0.627	0.599	0.613
	zero-shot	Flash 2.5	0.724	0.859	0.786
		27b	0.635	0.729	0.679
		12b	0.579	0.661	0.617
		4b	0.516	0.650	0.575
Español	one-shot	Flash 2.5	0.752	0.893	0.817
		27b	0.785	0.825	0.804
		12b	0.594	0.695	0.641
		4b	0.590	0.627	0.608
	zero-shot	Flash 2.5	0.711	0.819	0.761
		27b	0.635	0.729	0.679
		12b	0.641	0.667	0.654
		4b	0.496	0.672	0.571

Tab. 5.6: Resultados para *FECHA*.

Idioma	Estrategia	Modelo	Precision	Recall	F1-score
Inglés	one-shot	2.5 Flash	0.738	0.823	0.778
		27b	0.798	0.699	0.745
		12b	0.693	0.627	0.658
		4b	0.555	0.560	0.557
	zero-shot	2.5 Flash	0.703	0.804	0.750
		27b	0.694	0.684	0.689
		12b	0.642	0.593	0.617
		4b	0.567	0.565	0.566
Español	one-shot	2.5 Flash	0.713	0.818	0.762
		27b	0.819	0.737	0.776
		12b	0.713	0.617	0.662
		4b	0.577	0.612	0.594
	zero-shot	2.5 Flash	0.732	0.785	0.758
		27b	0.728	0.742	0.735
		12b	0.660	0.612	0.635
		4b	0.580	0.627	0.602

Tab. 5.7: Resultados para *ORGANIZACIÓN*.

5.2. Extracción de eventos

En la Tabla 5.8 se muestran los resultados obtenidos para cada prompt y modelo en la tarea de extracción de eventos. Se puede ver nuevamente cómo el rendimiento en la tarea de extracción aumenta con el tamaño del modelo.

La inclusión del campo *descripción* parece tener un efecto negativo en el desempeño de los modelos Gemma 3, ya que el valor de F1 disminuye en casi todos los campos. La única excepción es el campo *victimarios*, donde se observa una leve mejora al utilizar el modelo de mayor tamaño. Por otro lado, para el modelo Gemini 2.5 Flash, la adición de la descripción resulta en una mejora del F1 en todos los campos excepto *víctimas*, aunque esta mejora es muy leve.

En todos los casos, los modelos muestran mayor F1 en los campos *fecha* y *víctimas*, y en general, un F1 más bajo en el campo *victimarios*. Esto podría estar relacionado con la forma en que se menciona la información en los documentos, que suele ser más directa para las víctimas y las fechas. Además, la definición de *victimarios* incluye tanto autores intelectuales como materiales, lo cual amplía el espectro de posibles menciones y podría dificultar su identificación precisa.

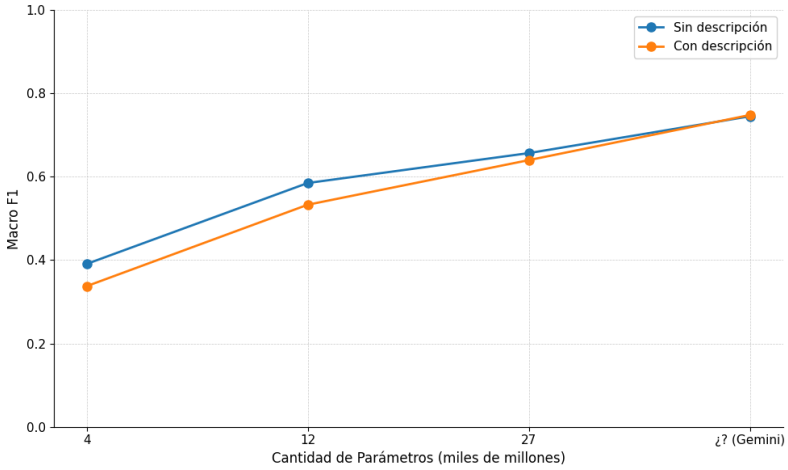

Fig. 5.2: Comparación del F1 Macro Average entre modelos y configuraciones para extracción de eventos.

Prompt	Modelo	Campo	Precisión	Recall	F1-Score
SIN DESCRIPCIÓN	4b	fecha	0.3212	0.4286	0.3683
		lugar	0.3214	0.4286	0.3673
		victimas	0.3914	0.5595	0.4606
		victimarios	0.3304	0.4128	0.3670
		totales	0.3412	0.4574	0.3908
	12b	fecha	0.5852	0.6270	0.6054
		lugar	0.5556	0.5952	0.5747
		victimas	0.6043	0.6601	0.6310
		victimarios	0.5272	0.5291	0.5281
		totales	0.5681	0.6028	0.5849
	27b	fecha	0.7179	0.6667	0.6914
		lugar	0.6325	0.5873	0.6091
		victimas	0.7429	0.7077	0.7248
		victimarios	0.6296	0.5747	0.6009
		totales	0.6807	0.6341	0.6566
	Gemini 2.5 Flash	fecha	0.7836	0.8333	0.8077
		lugar	0.7090	0.7540	0.7308
		victimas	0.7593	0.8056	0.7818
		victimarios	0.6391	0.6807	0.6592
		totales	0.7227	0.7684	0.7449
CON DESCRIPCIÓN	4b	fecha	0.3048	0.5079	0.3810
		lugar	0.2571	0.4286	0.3214
		victimas	0.2704	0.5053	0.3523
		victimarios	0.2402	0.3803	0.2944
		totales	0.2681	0.4555	0.3375
	12b	fecha	0.5105	0.5794	0.5428
		lugar	0.4895	0.5556	0.5204
		victimas	0.5322	0.6310	0.5774
		victimarios	0.4709	0.5114	0.4903
		totales	0.5008	0.5693	0.5328
	27b	fecha	0.6744	0.6905	0.6824
		lugar	0.5659	0.5794	0.5725
		victimas	0.6680	0.7196	0.6928
		victimarios	0.6137	0.6078	0.6107
		totales	0.6305	0.6493	0.6398
	Gemini 2.5 Flash	fecha	0.7910	0.8413	0.8154
		lugar	0.7239	0.7698	0.7462
		victimas	0.7264	0.7778	0.7512
		victimarios	0.6622	0.6974	0.6793
		totales	0.7259	0.7716	0.7480

Tab. 5.8: Resultados para cada modelo utilizando prompts con y sin descripción.

6. CONCLUSIONES Y TRABAJO FUTURO

6.1. Conclusiones

Los resultados obtenidos demuestran que, incluso con modelos de tamaño moderado ejecutados localmente y sin reentrenamiento, es posible alcanzar un rendimiento prometedor en la extracción de información.

Se observó una clara tendencia de mejora con el incremento del tamaño del modelo, siendo Gemini 2.5 Flash el que ofreció los mejores resultados generales en términos de precisión, recall y F1-score. En la tarea de extracción de entidades, Gemma 3 27b obtuvo resultados comparables, mientras que en extracción de eventos se impuso por amplio margen Gemini 2.5 Flash. Estos resultados sugieren que, para tareas de mayor complejidad, puede resultar conveniente utilizar modelos de mayor capacidad.

Asimismo, la estrategia de aprendizaje in-context one-shot, que incorpora un único ejemplo en el prompt, superó de manera general al enfoque zero-shot, subrayando el valor de proporcionar al modelo demostraciones concretas de la tarea. La ligera ventaja observada con prompts en español evidencia una buena adaptación de los modelos al idioma del corpus.

6.2. Trabajo futuro

Se identifican, en este breve apartado, algunas áreas de mejora que permitirán profundizar y optimizar los resultados obtenidos.

En primer lugar, vale la pena mencionar que la idea original fue trabajar con los datos de Abuelas de Plaza de Mayo, y en ese sentido, se propone continuar los experimentos con este conjunto de datos, además de definir una ontología específica que permita formalizar las entidades, relaciones y eventos relevantes para este dominio, facilitando así una extracción de información más precisa y estructurada.

Por otro lado, la superioridad observada en el enfoque one-shot indica que la incorporación de más ejemplos en el prompt podría potenciar significativamente el rendimiento de los modelos, por lo que resulta pertinente explorar esta variante en futuras experimentaciones.

Asimismo, se plantea refinar las instrucciones utilizadas en los prompts, resolviendo posibles ambigüedades y añadiendo reglas más específicas que orienten al modelo de manera más efectiva en la resolución de las tareas propuestas.

En cuanto a la evaluación, sería conveniente adoptar criterios más flexibles para considerar ciertos errores como aceptables, especialmente en contextos aplicados. Esto incluiría, por ejemplo, variaciones menores en la expresión de una entidad, segmentaciones alternativas o equivalencias semánticas que, sin alterar el sentido, podrían aceptarse según la finalidad práctica del sistema.

Además, se propone abordar aspectos no contemplados en esta etapa, como la resolución de correferencias (qué expresiones hacen referencia a la misma entidad), la homogeneización del formato de fechas y lugares, y el tratamiento de expresiones relativas (por ejemplo, “ayer” o “el mismo año”), con el fin de mejorar la coherencia y calidad de la información extraída.

Finalmente, se plantea extender el alcance del sistema hacia la extracción de relaciones entre entidades, lo que permitiría enriquecer las capacidades del modelo y ampliar las posibilidades de análisis e interpretación de los datos obtenidos.

7. ANEXO

Prompt 1: Extracción de Entidades (Español, Zero-shot)

```
<s>[INST]
Tu tarea es extraer todas las entidades relevantes del texto
    proporcionado. No omitas ninguna mención.
Las entidades deben ser de los siguientes tipos: PERSONA, ORGANIZACIÓN,
    LUGAR, FECHA.

**Nota importante:**
Extraer TODAS las menciones de la misma entidad como estén escritas.
Cualquier referencia a un día, fecha, momento temporal o periodo debe
    clasificarse como **FECHA**. Esto incluye nombres de días, meses,
    años, fechas completas o parciales, expresiones como "ayer", "el
    próximo martes", "en 1994", "hace dos semanas", etc.
Las edades NO deben clasificarse como FECHA.
Si una entidad tiene alias, inclúyelos como otra entidad aparte.
Para el tipo PERSONA, no incluir profesiones, cargos o títulos, solo
    nombres propios.
Para el tipo ORGANIZACIÓN, incluir nombres de instituciones, empresas,
    grupos políticos, etc.
Para el tipo ORGANIZACIÓN, extraer siglas y nombre de la organización por
    separado.
Para el tipo FECHA, extraer el fragmento más específico posible.
Si se menciona un grupo de nombres de lugares o instituciones enumerados
    juntos (por ejemplo: "las calles Fuerte Grande, Mineto y Funesti"),
    extraerlos en conjunto (la salida debe ser: "calles Fuerte Grande,
    Mineto y Funesti").
Extraer direcciones (calle y número), barrios, ciudades, etc. por separado.

Formato de salida ESTRICTO: JSON con una clave "entidades" que contenga
    una lista de objetos, donde cada objeto tiene "texto" y "tipo".
**ASEGÚRATE DE QUE EL JSON SEA VÁLIDO.** No incluyas comentarios,
    explicaciones ni ningún texto fuera del bloque JSON.
Ahora, procesa el siguiente texto y genera el JSON correspondiente:

Texto:
"{texto_entrada}"

JSON:
[/INST]"
```

Prompt 7.1: Extracción de Entidades (Español, Zero-shot).

Prompt 2: Extracción de Entidades (Inglés, Zero-shot)

```
"<s>[INST]
Your task is to extract all relevant entities from the provided text. Do
    not miss any mentions.
The entities must be of the following types: PERSON, ORGANIZATION,
    LOCATION, DATE.
```

```
**Important note:**
Extract ALL mentions of the same entity exactly as they appear.
Any reference to a day, date, point in time, or time period must be
    classified as a DATE. This includes names of days, months, years,
    complete or partial dates, and expressions like "yesterday," "next
    Tuesday," "in 1994," "two weeks ago," etc.
Ages should NOT be classified as a DATE.
If an entity has an alias, include it as a separate entity.
For the PERSON type, do not include professions, positions, or titles
    only proper names.
For the ORGANIZATION type, include the names of institutions, companies,
    political groups, etc.
For the ORGANIZATION type, extract acronyms and full organization names
    separately.
For the DATE type, extract the most specific fragment possible.
If a group of names of places or institutions is mentioned together (for
    example: "las calles Fuerte Grande, Mineto y Funesti"), extract them
    together as one entity (the output should be: "calles Fuerte Grande,
    Mineto y Funesti").
Extract addresses (street and number), neighborhoods, cities, etc., as
    separate entities.

STRICT output format: JSON with a key "entities" containing a list of
    objects, where each object has "text" and "type".
**MAKE SURE THE JSON IS VALID**. Do not include comments, explanations, or
    any text outside the JSON block.

Now, process the following text and generate the corresponding JSON:

Text:
"{texto_entrada}"

JSON:
[/INST]"
```

Prompt 7.2: Extracción de Entidades (Inglés, Zero-shot).

Prompt 3: Extracción de Entidades (Español, One-shot)

```
"<s>[INST]
Tu tarea es extraer todas las entidades relevantes del texto
    proporcionado. No omitas ninguna mención.
Las entidades deben ser de los siguientes tipos: PERSONA, ORGANIZACIÓN,
    LUGAR, FECHA.

**Nota importante:**
Extraer TODAS las menciones de la misma entidad como estén escritas.
Cualquier referencia a un día, fecha, momento temporal o periodo debe
    clasificarse como **FECHA**. Esto incluye nombres de días, meses,
    años, fechas completas o parciales, expresiones como "ayer", "el
    próximo martes", "en 1994", "hace dos semanas", etc.
Las edades NO deben clasificarse como FECHA.
Si una entidad tiene alias, inclúyelos como otra entidad aparte.
Para el tipo PERSONA, no incluir profesiones, cargos o títulos, solo
    nombres propios.
Para el tipo ORGANIZACIÓN, incluir nombres de instituciones, empresas,
    grupos políticos, etc.
```

Para el tipo ORGANIZACIÓN, extraer siglas y nombre de la organización por separado.
Para el tipo FECHA, extraer el fragmento más específico posible.
Si se menciona un grupo de nombres de lugares o instituciones enumerados juntos (por ejemplo: "las calles Fuerte Grande, Mineto y Funesti"), extraerlos en conjunto (la salida debe ser: "calles Fuerte Grande, Mineto y Funesti").
Extraer direcciones (calle y número), barrios, ciudades, etc. por separado.

Formato de salida ESTRICTO: JSON con una clave "entidades" que contenga una lista de objetos, donde cada objeto tiene "texto" y "tipo".
ASEGÚRATE DE QUE EL JSON SEA VÁLIDO. No incluyas comentarios, explicaciones ni ningún texto fuera del bloque JSON.

Ejemplo:
Texto de Ejemplo: "RESULTADO DEL INTERROGATORIO A PERSONAS AFINES
A LUCIO CABAÑAS BARRIENTOS

A las 7.00 horas del día de la fecha llegaron al Cam-
po Militar número Uno, nueve personas detenidas por la 27/a. Zo
na Militar, con sede en Acapulco, Gro., mismas que desde hace
dos meses se encontraban detenidas por sospechar que pertene---
cían al grupo de LUCIO CABAÑAS BARRIENTOS.

Los detenidos son: ALBERTO ARROYO DIONISIO, JUSTINO
BARRIENTOS, ROMANA RIOS DE ROQUE, DAVED ROJAS ARIAS, PETRONILO-
CASTRO HERNANDEZ, GUADALUPE CASTRO MOLINA, ISABEL JIMENEZ HERNAN
DEZ, Y LUIS CABAÑAS OCAMPO.

Agentes de esta Dirección procedieron de inmediato a
interrogar a las mencionadas personas, quienes han manifestado
lo siguiente:

LUIS CABAÑAS OCAMPO, dijo ser haber nacido en la calle San Luis, Corral Fal
so, Gro., casado, de 48 años de edad, agricultor, con domicilio-
conocido en San Vicente Benítez; ser tío de LUCIO CABAÑAS BAR
RRIENTOS y saber que éste se viene dedicando a actividades sub--
versivas en contra de los Gobiernos Estatal y Federal y que uni-
camente lo liga el parentesco que con éste tiene. Que fue miem--
bro del Frente Electoral del Pueblo (FEP) y de la U.E.N., ya que su ideología es iz--
quierdista, pero que no cree que la forma de liberar al pueblo
sea mediante la lucha armada. Que está dedicado a la labor del
campo y que en algunas ocasiones ha protestado por medio de la
televisión y mediante comunicados de prensa por las arbitrarie--
dades que comete el Ejército en los poblados de las Sierras de
Guerrero, cuando realizan la labor de ubicación de su pariente
LUCIO CABAÑAS; que no participa en dicho grupo y que poco tiempo
antes de ser detenido se entrevistó con el C. Gobernador del Es-
tado, quien le ofreció su intervención para que LUCIO CABAÑAS al
canzara la amnistía, entrevista que no logró realizar con este-
último porque al poco tiempo fue detenido por el secuestro que
realizó su pariente en contra de CUAUHTEMOC GARCIA TERAN y forzo
samente las autoridades de Guerrero han querido involucrarlo en-
dicho acto."

JSON de Ejemplo Esperado:
```json
```

```
{{
  "entidades": [
    {{"texto": "LUCIO CABAÑAS BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "7.00 horas del día de la fecha", "tipo": "FECHA"}},
    {{"texto": "Campo Militar número Uno", "tipo": "LUGAR"}},
    {{"texto": "27/a. Zona Militar", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Acapulco", "tipo": "LUGAR"}},
    {{"texto": "Gro.", "tipo": "LUGAR"}},
    {{"texto": "hace dos meses", "tipo": "FECHA"}},
    {{"texto": "LUCIO CABAÑAS BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "ALBERTO ARROYO DIONISIO", "tipo": "PERSONA"}},
    {{"texto": "JUSTINO BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "ROMANA RIOS DE ROQUE", "tipo": "PERSONA"}},
    {{"texto": "DAVID ROJAS ARIAS", "tipo": "PERSONA"}},
    {{"texto": "PETRONILO CASTRO HERNANDEZ", "tipo": "PERSONA"}},
    {{"texto": "GUADALUPE CASTRO MOLINA", "tipo": "PERSONA"}},
    {{"texto": "ISABEL JIMENEZ HERNANDEZ", "tipo": "PERSONA"}},
    {{"texto": "LUIS CABAÑAS OCAMPO", "tipo": "PERSONA"}},
    {{"texto": "LUIS CABAÑAS OCAMPO", "tipo": "PERSONA"}},
    {{"texto": "calle San Luis", "tipo": "LUGAR"}},
    {{"texto": "Corral Falso", "tipo": "LUGAR"}},
    {{"texto": "San Vicente Benítez", "tipo": "LUGAR"}},
    {{"texto": "LUCIO CABAÑAS BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "Gobiernos Estatal y Federal", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Frente Electoral del Pueblo", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "FEP", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "U.E.N.", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Ejército", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Sierras de Guerrero", "tipo": "LUGAR"}},
    {{"texto": "LUCIO CABAÑAS", "tipo": "PERSONA"}},
    {{"texto": "Gobernador del Estado", "tipo": "PERSONA"}},
    {{"texto": "LUCIO CABAÑAS", "tipo": "PERSONA"}},
    {{"texto": "CUAUHTEMOC GARCIA TERAN", "tipo": "PERSONA"}},
    {{"texto": "autoridades de Guerrero", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Guerrero", "tipo": "LUGAR"}}
  ]
}}
‘‘‘

Ahora, procesa el siguiente texto y genera el JSON correspondiente:

Texto:
"{texto_entrada}"

JSON:
[/INST]"
```

Prompt 7.3: Extracción de Entidades (Español, One-shot).

Prompt 4: Extracción de Entidades (Inglés, One-shot)

```
"<s>[INST]
Your task is to extract all relevant entities from the provided text. Do
    not miss any mentions.
The entities must be of the following types: PERSON, ORGANIZATION,
    LOCATION, DATE.

**Important note:**
Extract ALL mentions of the same entity exactly as they appear.
Any reference to a day, date, point in time, or time period must be
    classified as a DATE. This includes names of days, months, years,
    complete or partial dates, and expressions like "yesterday," "next
    Tuesday," "in 1994," "two weeks ago," etc.
Ages should NOT be classified as a DATE.
If an entity has an alias, include it as a separate entity.
For the PERSON type, do not include professions, positions, or titles
    only proper names.
For the ORGANIZATION type, include the names of institutions, companies,
    political groups, etc.
For the ORGANIZATION type, extract acronyms and full organization names
    separately.
For the DATE type, extract the most specific fragment possible.
If a group of names of places or institutions is mentioned together (for
    example: "las calles Fuerte Grande, Mineto y Funesti"), extract them
    together as one entity (the output should be: "calles Fuerte Grande,
    Mineto y Funesti").
Extract addresses (street and number), neighborhoods, cities, etc., as
    separate entities.

STRICT output format: JSON with a key "entities" containing a list of
    objects, where each object has "text" and "type".
**MAKE SURE THE JSON IS VALID**. Do not include comments, explanations, or
    any text outside the JSON block.

Example:
Example Text: "RESULTADO DEL INTERROGATORIO A PERSONAS AFINES
A LUCIO CABAÑAS BARRIENTOS

A las 7.00 horas del día de la fecha llegaron al Cam-
po Militar número Uno, nueve personas detenidas por la 27/a. Zo
na Militar, con sede en Acapulco, Gro., mismas que desde hace
dos meses se encontraban detenidas por sospechar que pertene---
cían al grupo de LUCIO CABAÑAS BARRIENTOS.

Los detenidos son: ALBERTO ARROYO DIONISIO, JUSTINO
BARRIENTOS, ROMANA RIOS DE ROQUE, DAVED ROJAS ARIAS, PETRONILO-
CASTRO HERNANDEZ, GUADALUPE CASTRO MOLINA, ISABEL JIMENEZ HERNAN
DEZ, Y LUIS CABAÑAS OCAMPO.

Agentes de esta Dirección procedieron de inmediato a
interrogar a las mencionadas personas, quienes han manifestado
lo siguiente:

LUIS CABAÑAS OCAMPO, dijo ser haber nacido en la calle San Luis, Corral Fal
so, Gro., casado, de 48 años de edad, agricultor, con domicilio-
conocido en San Vicente Benítez; ser tío de LUCIO CABAÑAS BAR
RRIENTOS y saber que éste se viene dedicando a actividades sub--
```

```
versivas en contra de los Gobiernos Estatal y Federal y que uni-
camente lo liga el parentesco que con éste tiene. Que fue miem--
bro del Frente Electoral del Pueblo (FEP) y de la U.E.N., ya que su
    ideología es iz--
quierdista, pero que no cree que la forma de liberar al pueblo
sea mediante la lucha armada. Que está dedicado a la labor del
campo y que en algunas ocasiones ha protestado por medio de la
televisión y mediante comunicados de prensa por las arbitrarie--
dades que comete el Ejército en los poblados de las Sierras de
Guerrero, cuando realizan la labor de ubicación de su pariente
LUCIO CABAÑAS; que no participa en dicho grupo y que poco tiempo
antes de ser detenido se entrevistó con el C. Gobernador del Es-
tado, quien le ofreció su intervención para que LUCIO CABAÑAS al
canzara la amnistía, entrevista que no logró realizar con este-
último porque al poco tiempo fue detenido por el secuestro que
realizó su pariente en contra de CUAUHTEMOC GARCIA TERAN y forzo
samente las autoridades de Guerrero han querido involucrarlo en-
dicho acto."

Expected Example JSON:
‘‘‘json
{{
  "entities": [
    {{"text": "LUCIO CABAÑAS BARRIENTOS", "type": "PERSON"}},
    {{"text": "7.00 horas del día de la fecha", "type": "DATE"}},
    {{"text": "Campo Militar número Uno", "type": "LOCATION"}},
    {{"text": "27/a. Zona Militar", "type": "ORGANIZATION"}},
    {{"text": "Acapulco", "type": "LOCATION"}},
    {{"text": "Gro.", "type": "LOCATION"}},
    {{"text": "hace dos meses", "type": "DATE"}},
    {{"text": "LUCIO CABAÑAS BARRIENTOS", "type": "PERSON"}},
    {{"text": "ALBERTO ARROYO DIONISIO", "type": "PERSON"}},
    {{"text": "JUSTINO BARRIENTOS", "type": "PERSON"}},
    {{"text": "ROMANA RIOS DE ROQUE", "type": "PERSON"}},
    {{"text": "DAVID ROJAS ARIAS", "type": "PERSON"}},
    {{"text": "PETRONILO CASTRO HERNANDEZ", "type": "PERSON"}},
    {{"text": "GUADALUPE CASTRO MOLINA", "type": "PERSON"}},
    {{"text": "ISABEL JIMENEZ HERNANDEZ", "type": "PERSON"}},
    {{"text": "LUIS CABAÑAS OCAMPO", "type": "PERSON"}},
    {{"text": "LUIS CABAÑAS OCAMPO", "type": "PERSON"}},
    {{"texto": "calle San Luis", "tipo": "LUGAR"}},
    {{"text": "Corral Falso", "type": "LOCATION"}},
    {{"text": "San Vicente Benítez", "type": "LOCATION"}},
    {{"text": "LUCIO CABAÑAS BARRIENTOS", "type": "PERSON"}},
    {{"text": "Gobiernos Estatal y Federal", "type": "ORGANIZATION"}},
    {{"text": "Frente Electoral del Pueblo", "type": "ORGANIZATION"}},
    {{"text": "FEP", "type": "ORGANIZATION"}},
    {{"text": "U.E.N.", "type": "ORGANIZATION"}},
    {{"text": "Ejército", "type": "ORGANIZATION"}},
    {{"text": "Sierras de Guerrero", "type": "LOCATION"}},
    {{"text": "LUCIO CABAÑAS", "type": "PERSON"}},
    {{"text": "Gobernador del Estado", "type": "PERSON"}},
    {{"text": "LUCIO CABAÑAS", "type": "PERSON"}},
    {{"text": "CUAUHTEMOC GARCIA TERAN", "type": "PERSON"}},
    {{"text": "autoridades de Guerrero", "type": "ORGANIZATION"}},
    {{"text": "Guerrero", "type": "LOCATION"}}
  ]
}}
```

```
‘‘‘

Now, process the following text and generate the corresponding JSON:

Text:
"{texto_entrada}"

JSON:
[/INST]"
```

Prompt 7.4: Extracción de Entidades (Inglés, One-shot).

Prompt 5: Extracción de eventos sin descripción

```
"<s>[INST]
Tu tarea es identificar **eventos** en el texto original, basándote en la
    ontología de tipos de evento definida a continuación.

**Tipos de Evento Permitidos**:
- CAPTURA: secuestro, captura, detención o desaparición de personas.
- ASESINATO: asesinato de personas.


**Cada evento debe contener los siguientes campos**:
- "tipo": Uno de los valores permitidos (CAPTURA, ASESINATO).
- "fecha": Valor textual que coincida exactamente con el texto y que
    corresponda a tipo FECHA (horas, días, meses, años, rangos o periodos
    de tiempo, incluyendo relativos como "día de la fecha"). Si no se
    menciona, usa ‘null‘.
- "lugar": Valor textual que coincida exactamente con el texto y que
    corresponda a tipo LUGAR (puede ser relativo o general). Si no se
    menciona, usa ‘null‘.
- "victimas": Lista de valores textuales de las entidades de tipo PERSONA
    afectadas por el evento. Definición de victima: Persona que sufre
    directamente las consecuencias del evento (captura o asesinato).
    - Si se conoce el nombre de la persona, úsalo.
    - Si se menciona una organización o grupo armado, extrae el nombre de
    la organización directamente y omite expresiones como "miembros de",
    "adeptos de", "integrantes de", etc.
    - Si no se conoce el nombre de la persona pero sí su organización o
    grupo, incluye el nombre de la organización o grupo (una vez por
    mención de grupo). Incluir esto sólo cuando no se conozca el nombre de
    la persona.
    - Si no se conoce el nombre de la persona ni su organización, pero se
    indica una cantidad o profesión (ej. "dos campesinos", "varios
    individuos"), utiliza esa descripción.
    - Si no se aporta información específica más allá de una existencia
    genérica, utiliza "otro" (singular) u "otros" (plural).
    - Si no se menciona explícitamente ninguna víctima, la lista debe ser
    vacía ‘[]‘.
- "victimarios": Lista de valores textuales de las entidades de tipo
    PERSONA u ORGANIZACIÓN responsables del evento (incluir presuntos
    responsables). Definición de victimario: Persona u organización que
    causa o ejecuta el evento (captura o asesinato). Incluye autores
    materiales y quienes dan la orden.
    - Aplican las mismas reglas que para "victimas" (nombre de persona,
    organización, descripción genérica, "otro/s").
```

```
    - Si no se menciona explícitamente ningún victimario, la lista debe
    ser vacía `[]`.

**Instrucciones Estrictas**:
1.  Debes identificar y extraer **TODOS** los eventos de los tipos
    permitidos (CAPTURA, ASESINATO) que se mencionen explícitamente en el
    texto, **EN EL ORDEN** en que aparecen. No omitas ninguno.
2.  Los eventos pueden ser hechos consumados, planeados o en proceso de
    ejecución.
3.  ES ABSOLUTAMENTE CRÍTICO que los valores para el campo "tipo"
    provengan ÚNICAMENTE de la lista de "Tipos de Evento Permitidos". Si
    un evento no encaja estrictamente en esas categorías, NO lo incluyas.
    NO incluyas asaltos, robos, ataques, enfrentamientos, emboscadas, ni
    ningún otro tipo de evento que no esté explícitamente en la lista de
    tipos permitidos.
4.  Solo extrae eventos que sean mencionados directa y explícitamente en
    el texto original. No infieras eventos no descritos.
5.  Si para un evento no se menciona explícitamente información para los
    campos "fecha", "lugar", la lista de "victimas" está vacía o la lista
    de "victimarios" está vacía, usa `null` para "fecha" y "lugar", y
    listas vacías `[]` para "victimas" y "victimarios" respectivamente.
6.  **Regla especial para CAPTURA**: Si se menciona que una persona
    "presta declaración", "es interrogada" o "está detenida" ante una
    autoridad, y no se ha mencionado previamente un evento de captura
    explícito para esa persona en ese contexto, considera esto como un
    evento de tipo CAPTURA. La autoridad interrogadora sería el victimario.
7.  En los eventos de tipo CAPTURA, si hay más de una mención de LUGAR,
    incluir sólo el LUGAR inicial.
8.  Extraer una única mención de cada PERSONA.
9.  **MUY IMPORTANTE:**Si no hay eventos válidos que cumplan con los
    criterios en el texto, devuelve un JSON con una clave "eventos" que
    contenga una lista vacía: `{{"eventos": []}}`.

**Formato de salida ESTRICTO**:
Un único objeto JSON con una clave "eventos". El valor de "eventos" debe
    ser una lista de objetos. Cada objeto de la lista representa un evento
    y debe tener las claves "tipo", "fecha", "lugar", "victimas" y
    "victimarios".

**Ejemplo de Texto**:
"RESULTADO DEL INTERROGATORIO A PERSONAS AFINES
A LUCIO CABAÑAS BARRIENTOS

A las 7.00 horas del día de la fecha llegaron al Cam-
po Militar número Uno, nueve personas detenidas por la 27/a. Zo
na Militar, con sede en Acapulco, Gro., mismas que desde hace
dos meses se encontraban detenidas por sospechar que pertene---
cían al grupo de LUCIO CABAÑAS BARRIENTOS.

Los detenidos son: ALBERTO ARROYO DIONISIO, JUSTINO
BARRIENTOS, ROMANA RIOS DE ROQUE, DAVED ROJAS ARIAS, PETRONILO-
CASTRO HERNANDEZ, GUADALUPE CASTRO MOLINA, ISABEL JIMENEZ HERNAN
DEZ, Y LUIS CABAÑAS OCAMPO.

Agentes de esta Dirección procedieron de inmediato a
interrogar a las mencionadas personas, quienes han manifestado
lo siguiente:
```

LUIS CABAÑAS OCAMPO, dijo ser haber nacido en la calle San Luis, Corral Fal
so, Gro., casado, de 48 años de edad, agricultor, con domicilio-
conocido en San Vicente Benítez; ser tío de LUCIO CABAÑAS BAR
RRIENTOS y saber que éste se viene dedicando a actividades sub--
versivas en contra de los Gobiernos Estatal y Federal y que uni-
camente lo liga el parentesco que con éste tiene. Que fue miem--
bro del Frente Electoral del Pueblo (FEP) y de la U.E.N., ya que su
ideología es iz--
quierdista, pero que no cree que la forma de liberar al pueblo
sea mediante la lucha armada. Que está dedicado a la labor del
campo y que en algunas ocasiones ha protestado por medio de la
televisión y mediante comunicados de prensa por las arbitrarie--
dades que comete el Ejército en los poblados de las Sierras de
Guerrero, cuando realizan la labor de ubicación de su pariente
LUCIO CABAÑAS; que no participa en dicho grupo y que poco tiempo
antes de ser detenido se entrevistó con el C. Gobernador del Es-
tado, quien le ofreció su intervención para que LUCIO CABAÑAS al
canzara la amnistía, entrevista que no logró realizar con este-
último porque al poco tiempo fue detenido por el secuestro que
realizó su pariente en compañia de dos compañeros en contra de CUAUHTEMOC
GARCIA TERAN y forzo
samente las autoridades de Guerrero han querido involucrarlo en-
dicho acto.

14.DIC-79

[Firma]
BARRIENTOS FLORES JUSTINO

Se tiene conocimiento que se incorporó al Partido
de los Pobres que comandaba Lucio Cabañas Barrientos, desde
principios del año de 1972.
90
Con motivo de que Lucio Cabañas Barrientos se sen--
tía acosado por elementos de la fuerza pública en la Sierra
de Guerrero y ante el temor de sus adeptos de ser capturados-
que abandonaban su causa, organizó un grupo de sujetos a quie
nes encomendó la misión específica de obligar a los deserto--
res a reincorporarse al grupo que comandaba él mismo.

Por lo anterior se sabe que a 5 días de su vuelta,
fue sustraido en forma violenta de su domicilio en Atoyac de-
Alvarez, Justino Barrientos Flores y obligado a participar en
diferentes hechos delictuosos.

El 25 de junio de 1972, resultó muerto en una embos
cada que realizaron en contra de elementos del Ejército mexi-
cano del 50/o Batallón de Infantería quienes se trasladaban a
su base de partida en San Vicente de Benitez, Gro., mismos
que al sentirse atacados repelieron la agresión dando como re
sultado la muerte de varios individuos entre los que se encon
traba este sujeto."

Entidades ya identificadas en el texto original (Ejemplo):
```json
{{
  "entidades": [
    {{"texto": "LUCIO CABAÑAS BARRIENTOS", "tipo": "PERSONA"}},
```

```
    {{"texto": "7.00 horas del día de la fecha", "tipo": "FECHA"}},
    {{"texto": "Campo Militar número Uno", "tipo": "LUGAR"}},
    {{"texto": "27/a. Zona Militar", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Acapulco. Gro.", "tipo": "LUGAR"}},
    {{"texto": "hace dos meses", "tipo": "FECHA"}},
    {{"texto": "LUCIO CABAÑAS BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "ALBERTO ARROYO DIONISIO", "tipo": "PERSONA"}},
    {{"texto": "JUSTINO BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "ROMANA RIOS DE ROQUE", "tipo": "PERSONA"}},
    {{"texto": "DAVED ROJAS ARIAS", "tipo": "PERSONA"}},
    {{"texto": "PETRONILO CASTRO HERNANDEZ", "tipo": "PERSONA"}},
    {{"texto": "GUADALUPE CASTRO MOLINA", "tipo": "PERSONA"}},
    {{"texto": "ISABEL JIMENEZ HERNANDEZ", "tipo": "PERSONA"}},
    {{"texto": "LUIS CABAÑAS OCAMPO", "tipo": "PERSONA"}},
    {{"texto": "LUIS CABAÑAS OCAMPO", "tipo": "PERSONA"}},
    {{"texto": "calle San Luis. Corral Falso, Gro.", "tipo": "LUGAR"}},
    {{"texto": "San Vicente Benítez", "tipo": "LUGAR"}},
    {{"texto": "LUCIO CABAÑAS BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "Gobiernos Estatal y Federal", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Frente Electoral del Pueblo", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "FEP", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "U.E.N.", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Ejército", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Sierras de Guerrero", "tipo": "LUGAR"}},
    {{"texto": "LUCIO CABAÑAS", "tipo": "PERSONA"}},
    {{"texto": "Gobernador del Estado", "tipo": "PERSONA"}},
    {{"texto": "LUCIO CABAÑAS", "tipo": "PERSONA"}},
    {{"texto": "CUAUHTEMOC GARCIA TERAN", "tipo": "PERSONA"}},
    {{"texto": "autoridades de Guerrero", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Guerrero", "tipo": "LUGAR"}},
    {{"texto": "14.DIC-79", "tipo": "FECHA"}},
    {{"texto": "BARRIENTOS FLORES JUSTINO", "tipo": "PERSONA"}},
    {{"texto": "Partido de los Pobres", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Lucio Cabañas Barrientos", "tipo": "PERSONA"}},
    {{"texto": "principios del año de 1972", "tipo": "FECHA"}},
    {{"texto": "Lucio Cabañas Barrientos", "tipo": "PERSONA"}},
    {{"texto": "Sierra de Guerrero", "tipo": "LUGAR"}},
    {{"texto": "a 5 días de su vuelta", "tipo": "FECHA"}},
    {{"texto": "su domicilio en Atoyac de Alvarez", "tipo": "LUGAR"}},
    {{"texto": "Justino Barrientos Flores", "tipo": "PERSONA"}},
    {{"texto": "25 de junio de 1972", "tipo": "FECHA"}},
    {{"texto": "Ejército mexi-cano", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "50/o Batallón de Infantería", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "San Vicente de Benitez, Gro.", "tipo": "LUGAR"}},
  ]
}}
```

JSON de ejemplo esperado:

```json
{{
  "eventos": [
    {{
      "tipo": "CAPTURA",
      "fecha": "7.00 horas del día de la fecha",
      "lugar": "Campo Militar número Uno",
      "victimas": [
        "ALBERTO ARROYO DIONISIO",
```

```
        "JUSTINO BARRIENTOS",
        "ROMANA RIOS DE ROQUE",
        "DAVED ROJAS ARIAS",
        "PETRONILO CASTRO HERNANDEZ",
        "GUADALUPE CASTRO MOLINA",
        "ISABEL JIMENEZ HERNANDEZ",
        "LUIS CABAÑAS OCAMPO"
      ],
      "victimarios": ["27/a. Zona Militar"]
    }},
    {{
      "tipo": "CAPTURA",
      "fecha": "poco tiempo antes",
      "lugar": null,
      "victimas": [
        "LUIS CABAÑAS OCAMPO"
      ],
      "victimarios": []
    }},
    {{
      "tipo": "CAPTURA",
      "fecha": null,
      "lugar": null,
      "victimas": [
        "CUAUHTEMOC GARCIA TERAN"
      ],
      "victimarios": [
        "LUCIO CABAÑAS BARRIENTOS",
        "dos compañeros"
      ]
    }},
    {{
      "tipo": "CAPTURA",
      "fecha": "a 5 días de su vuelta",
      "lugar": "su domicilio en Atoyac de-Alvarez",
      "victimas": [
        "Justino Barrientos Flores"
      ],
      "victimarios": []
    }},
    {{
      "tipo": "ASESINATO",
      "fecha": "25 de junio de 1972",
      "lugar": "San Vicente de Benitez, Gro.",
      "victimas": [
        "Justino Barrientos Flores", "varios individuos"
      ],
      "victimarios": [
        "Ejército mexicano", "50/o Batallón de Infantería"
      ]
    }}
  ]
}}
```

Utiliza las entidades previamente extraídas como base para completar los campos.

```
Ahora, procesa el siguiente texto y genera el JSON correspondiente:

**Texto**:
"{texto_entrada}"

Entidades ya identificadas en el texto original:
```json
{entidades_extraidas_str}
```

JSON:
[/INST]"
```

Prompt 7.5: Extracción de eventos sin descripción.

Prompt 6: Extracción de eventos con descripción

```
"<s>[INST]
Tu tarea es identificar **eventos** en el texto original, basándote en la
    ontología de tipos de evento definida a continuación.

**Tipos de Evento Permitidos**:
- CAPTURA: secuestro, captura, detención o desaparición de personas.
- ASESINATO: asesinato de personas.


**Cada evento debe contener los siguientes campos**:
- "descripcion": Debe ser la frase literal o el conjunto de frases
    contiguas que describen el evento en el texto original, sin añadir
    contexto adicional ni información no relacionada con el evento
    específico. Siempre debe tener un valor y debe describir un evento
    dentro de los tipos permitidos.
- "tipo": Uno de los valores permitidos (CAPTURA, ASESINATO).
- "fecha": Valor textual que coincida exactamente con el texto y que
    corresponda a tipo FECHA (horas, días, meses, años, rangos o periodos
    de tiempo, incluyendo relativos como "día de la fecha"). Si no se
    menciona, usa `null`.
- "lugar": Valor textual que coincida exactamente con el texto y que
    corresponda a tipo LUGAR (puede ser relativo o general). Si no se
    menciona, usa `null`.
- "victimas": Lista de valores textuales de las entidades de tipo PERSONA
    afectadas por el evento. Definición de victima: Persona que sufre
    directamente las consecuencias del evento (captura o asesinato).
    - Si se conoce el nombre de la persona, úsalo.
    - Si se menciona una organización o grupo armado, extrae el nombre de
    la organización directamente y omite expresiones como "miembros de",
    "adeptos de", "integrantes de", etc.
    - Si no se conoce el nombre de la persona pero sí su organización o
    grupo, incluye el nombre de la organización o grupo (una vez por
    mención de grupo). Incluir esto sólo cuando no se conozca el nombre de
    la persona.
    - Si no se conoce el nombre de la persona ni su organización, pero se
    indica una cantidad o profesión (ej. "dos campesinos", "varios
    individuos"), utiliza esa descripción.
    - Si no se aporta información específica más allá de una existencia
    genérica, utiliza "otro" (singular) u "otros" (plural).
```

```
    - Si no se menciona explícitamente ninguna víctima, la lista debe ser
    vacía `[]`.
- "victimarios": Lista de valores textuales de las entidades de tipo
    PERSONA u ORGANIZACIÓN responsables del evento (incluir presuntos
    responsables). Definición de victimario: Persona u organización que
    causa o ejecuta el evento (captura o asesinato). Incluye autores
    materiales y quienes dan la orden.
    - Aplican las mismas reglas que para "victimas" (nombre de persona,
    organización, descripción genérica, "otro/s").
    - Si no se menciona explícitamente ningún victimario, la lista debe
    ser vacía `[]`.

**Instrucciones Estrictas**:
1.  Debes identificar y extraer **TODOS** los eventos de los tipos
    permitidos (CAPTURA, ASESINATO) que se mencionen explícitamente en el
    texto, **EN EL ORDEN** en que aparecen. No omitas ninguno.
2.  Los eventos pueden ser hechos consumados, planeados o en proceso de
    ejecución.
3.  ES ABSOLUTAMENTE CRÍTICO que los valores para el campo "tipo"
    provengan ÚNICAMENTE de la lista de "Tipos de Evento Permitidos". Si
    un evento no encaja estrictamente en esas categorías, NO lo incluyas.
    NO incluyas asaltos, robos, ataques, enfrentamientos, emboscadas, ni
    ningún otro tipo de evento que no esté explícitamente en la lista de
    tipos permitidos.
4.  Solo extrae eventos que sean mencionados directa y explícitamente en
    el texto original. No infieras eventos no descritos.
5.  Si para un evento no se menciona explícitamente información para los
    campos "fecha", "lugar", la lista de "victimas" está vacía o la lista
    de "victimarios" está vacía, usa `null` para "fecha" y "lugar", y
    listas vacías `[]` para "victimas" y "victimarios respectivamente. El
    campo "descripcion" siempre debe tener contenido.
6.  **Regla especial para CAPTURA**: Si se menciona que una persona
    "presta declaración", "es interrogada" o "está detenida" ante una
    autoridad, y no se ha mencionado previamente un evento de captura
    explícito para esa persona en ese contexto, considera esto como un
    evento de tipo CAPTURA. La autoridad interrogadora sería el victimario.
7.  En los eventos de tipo CAPTURA, si hay más de una mención de LUGAR,
    incluir sólo el LUGAR inicial.
8.  Extraer una única mención de cada PERSONA.
9.  **MUY IMPORTANTE:**Si no hay eventos válidos que cumplan con los
    criterios en el texto, devuelve un JSON con una clave "eventos" que
    contenga una lista vacía: `{{"eventos": []}}`.

**Formato de salida ESTRICTO**:
Un único objeto JSON con una clave "eventos". El valor de "eventos" debe
    ser una lista de objetos. Cada objeto de la lista representa un evento
    y debe tener las claves "descripcion", "tipo", "fecha", "lugar",
    "victimas" y "victimarios".

**Ejemplo de Texto**:
"RESULTADO DEL INTERROGATORIO A PERSONAS AFINES
A LUCIO CABAÑAS BARRIENTOS

A las 7.00 horas del día de la fecha llegaron al Cam-
po Militar número Uno, nueve personas detenidas por la 27/a. Zo
na Militar, con sede en Acapulco, Gro., mismas que desde hace
dos meses se encontraban detenidas por sospechar que pertene---
cían al grupo de LUCIO CABAÑAS BARRIENTOS.
```

Los detenidos son: ALBERTO ARROYO DIONISIO, JUSTINO
BARRIENTOS, ROMANA RIOS DE ROQUE, DAVED ROJAS ARIAS, PETRONILO-
CASTRO HERNANDEZ, GUADALUPE CASTRO MOLINA, ISABEL JIMENEZ HERNAN
DEZ, Y LUIS CABAÑAS OCAMPO.

Agentes de esta Dirección procedieron de inmediato a
interrogar a las mencionadas personas, quienes han manifestado
lo siguiente:

LUIS CABAÑAS OCAMPO, dijo ser haber nacido en la calle San Luis, Corral Fal
so, Gro., casado, de 48 años de edad, agricultor, con domicilio-
conocido en San Vicente Benítez; ser tío de LUCIO CABAÑAS BAR
RRIENTOS y saber que éste se viene dedicando a actividades sub--
versivas en contra de los Gobiernos Estatal y Federal y que uni-
camente lo liga el parentesco que con éste tiene. Que fue miem--
bro del Frente Electoral del Pueblo (FEP) y de la U.E.N., ya que su
ideología es iz--
quierdista, pero que no cree que la forma de liberar al pueblo
sea mediante la lucha armada. Que está dedicado a la labor del
campo y que en algunas ocasiones ha protestado por medio de la
televisión y mediante comunicados de prensa por las arbitrarie--
dades que comete el Ejército en los poblados de las Sierras de
Guerrero, cuando realizan la labor de ubicación de su pariente
LUCIO CABAÑAS; que no participa en dicho grupo y que poco tiempo
antes de ser detenido se entrevistó con el C. Gobernador del Es-
tado, quien le ofreció su intervención para que LUCIO CABAÑAS al
canzara la amnistía, entrevista que no logró realizar con este-
último porque al poco tiempo fue detenido por el secuestro que
realizó su pariente en compañia de dos compañeros en contra de CUAUHTEMOC
GARCIA TERAN y forzo
samente las autoridades de Guerrero han querido involucrarlo en-
dicho acto.

14.DIC-79

[Firma]
BARRIENTOS FLORES JUSTINO

Se tiene conocimiento que se incorporó al Partido
de los Pobres que comandaba Lucio Cabañas Barrientos, desde
principios del año de 1972.
90
Con motivo de que Lucio Cabañas Barrientos se sen--
tía acosado por elementos de la fuerza pública en la Sierra
de Guerrero y ante el temor de sus adeptos de ser capturados-
que abandonaban su causa, organizó un grupo de sujetos a quie
nes encomendó la misión específica de obligar a los deserto--
res a reincorporarse al grupo que comandaba él mismo.

Por lo anterior se sabe que a 5 días de su vuelta,
fue sustraido en forma violenta de su domicilio en Atoyac de-
Alvarez, Justino Barrientos Flores y obligado a participar en
diferentes hechos delictuosos.

El 25 de junio de 1972, resultó muerto en una embos
cada que realizaron en contra de elementos del Ejército mexi-
cano del 50/o Batallón de Infantería quienes se trasladaban a

```
su base de partida en San Vicente de Benitez, Gro., mismos
que al sentirse atacados repelieron la agresión dando como re
sultado la muerte de varios individuos entre los que se encon
traba este sujeto."

**Entidades ya identificadas en el texto original (Ejemplo)**:
```json
{{
  "entidades": [
    {{"texto": "LUCIO CABAÑAS BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "7.00 horas del día de la fecha", "tipo": "FECHA"}},
    {{"texto": "Campo Militar número Uno", "tipo": "LUGAR"}},
    {{"texto": "27/a. Zona Militar", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Acapulco. Gro.", "tipo": "LUGAR"}},
    {{"texto": "hace dos meses", "tipo": "FECHA"}},
    {{"texto": "LUCIO CABAÑAS BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "ALBERTO ARROYO DIONISIO", "tipo": "PERSONA"}},
    {{"texto": "JUSTINO BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "ROMANA RIOS DE ROQUE", "tipo": "PERSONA"}},
    {{"texto": "DAVED ROJAS ARIAS", "tipo": "PERSONA"}},
    {{"texto": "PETRONILO CASTRO HERNANDEZ", "tipo": "PERSONA"}},
    {{"texto": "GUADALUPE CASTRO MOLINA", "tipo": "PERSONA"}},
    {{"texto": "ISABEL JIMENEZ HERNANDEZ", "tipo": "PERSONA"}},
    {{"texto": "LUIS CABAÑAS OCAMPO", "tipo": "PERSONA"}},
    {{"texto": "LUIS CABAÑAS OCAMPO", "tipo": "PERSONA"}},
    {{"texto": "calle San Luis. Corral Falso, Gro.", "tipo": "LUGAR"}},
    {{"texto": "San Vicente Benítez", "tipo": "LUGAR"}},
    {{"texto": "LUCIO CABAÑAS BARRIENTOS", "tipo": "PERSONA"}},
    {{"texto": "Gobiernos Estatal y Federal", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Frente Electoral del Pueblo", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "FEP", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "U.E.N.", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Ejército", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Sierras de Guerrero", "tipo": "LUGAR"}},
    {{"texto": "LUCIO CABAÑAS", "tipo": "PERSONA"}},
    {{"texto": "Gobernador del Estado", "tipo": "PERSONA"}},
    {{"texto": "LUCIO CABAÑAS", "tipo": "PERSONA"}},
    {{"texto": "CUAUHTEMOC GARCIA TERAN", "tipo": "PERSONA"}},
    {{"texto": "autoridades de Guerrero", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Guerrero", "tipo": "LUGAR"}},
    {{"texto": "14.DIC-79", "tipo": "FECHA"}},
    {{"texto": "BARRIENTOS FLORES JUSTINO", "tipo": "PERSONA"}},
    {{"texto": "Partido de los Pobres", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "Lucio Cabañas Barrientos", "tipo": "PERSONA"}},
    {{"texto": "principios del año de 1972", "tipo": "FECHA"}},
    {{"texto": "Lucio Cabañas Barrientos", "tipo": "PERSONA"}},
    {{"texto": "Sierra de Guerrero", "tipo": "LUGAR"}},
    {{"texto": "a 5 días de su vuelta", "tipo": "FECHA"}},
    {{"texto": "su domicilio en Atoyac de Alvarez", "tipo": "LUGAR"}},
    {{"texto": "Justino Barrientos Flores", "tipo": "PERSONA"}},
    {{"texto": "25 de junio de 1972", "tipo": "FECHA"}},
    {{"texto": "Ejército mexi-cano", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "50/o Batallón de Infantería", "tipo": "ORGANIZACIÓN"}},
    {{"texto": "San Vicente de Benitez, Gro.", "tipo": "LUGAR"}},
  ]
}}
```
```

```
**JSON de ejemplo esperado**:
‘‘‘json
{{
  "eventos": [
    {{
      "descripcion": "A las 7.00 horas del día de la fecha llegaron al
    Campo Militar número Uno, nueve personas detenidas por la 27/a. Zona
    Militar, con sede en Acapulco, Gro.",
      "tipo": "CAPTURA",
      "fecha": "7.00 horas del día de la fecha",
      "lugar": "Campo Militar número Uno",
      "victimas": [
        "ALBERTO ARROYO DIONISIO",
        "JUSTINO BARRIENTOS",
        "ROMANA RIOS DE ROQUE",
        "DAVED ROJAS ARIAS",
        "PETRONILO CASTRO HERNANDEZ",
        "GUADALUPE CASTRO MOLINA",
        "ISABEL JIMENEZ HERNANDEZ",
        "LUIS CABAÑAS OCAMPO"
      ],
      "victimarios": ["27/a. Zona Militar"]
    }},
    {{
      "descripcion": "poco tiempo antes de ser detenido",
      "tipo": "CAPTURA",
      "fecha": "poco tiempo antes",
      "lugar": null,
      "victimas": [
        "LUIS CABAÑAS OCAMPO"
      ],
      "victimarios": []
    }},
    {{
      "descripcion": "el secuestro que realizó su pariente en compañia de
    dos compañeros en contra de CUAUHTEMOC GARCIA TERAN",
      "tipo": "CAPTURA",
      "fecha": null,
      "lugar": null,
      "victimas": [
        "CUAUHTEMOC GARCIA TERAN"
      ],
      "victimarios": [
        "LUCIO CABAÑAS BARRIENTOS",
        "dos compañeros"
      ]
    }},
    {{
      "descripcion": "fue sustraido en forma violenta de su domicilio en
    Atoyac de-Alvarez, Justino Barrientos Flores",
      "tipo": "CAPTURA",
      "fecha": "a 5 días de su vuelta",
      "lugar": "su domicilio en Atoyac de-Alvarez",
      "victimas": [
        "Justino Barrientos Flores"
      ],
      "victimarios": []
    }},
```

```
    {{
      "descripcion": "El 25 de junio de 1972, resultó muerto en una
    emboscada que realizaron en contra de elementos del Ejército mexicano
    del 50/o Batallón de Infantería quienes se trasladaban a su base de
    partida en San Vicente de Benitez, Gro., mismos que al sentirse
    atacados repelieron la agresión dando como resultado la muerte de
    varios individuos entre los que se encontraba este sujeto.",
      "tipo": "ASESINATO",
      "fecha": "25 de junio de 1972",
      "lugar": "San Vicente de Benitez, Gro.",
      "victimas": [
        "Justino Barrientos Flores", "varios individuos"
      ],
      "victimarios": [
        "Ejército mexicano", "50/o Batallón de Infantería"
      ]
    }}
  ]
}}
```

```
Utiliza las entidades previamente extraídas como base para completar los
    campos.


Ahora, procesa el siguiente texto y genera el JSON correspondiente:

**Texto**:
"{texto_entrada}"

Entidades ya identificadas en el texto original:
```json
{entidades_extraidas_str}
```

JSON:
[/INST]"
```

Prompt 7.6: Extracción de eventos con descripción.

Bibliografía

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, 2025. Online manuscript released January 12, 2025.

[2] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey, 2024.

[3] Abuelas de Plaza de Mayo. Proyecto de inteligencia artificial para colaborar con la búsqueda de nietos y nietas. `https://www.abuelas.org.ar/prensa-y-difusion/noticias/2035`, 2025. Consultado el 10 de julio de 2025.

[4] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[9] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[11] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[12] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[13] Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. Multilegalpile: A 689gb multilingual legal corpus, 2024.

[14] Comisión de la Verdad. Modelo NER. `https://www.comisiondelaverdad.co/modelo-ner`, 2022. Último acceso: 7 de junio de 2025.

[15] Marianela Ciolfi Felice, Ivana Feldfeber, Carolina Glasserman Apicella, Yasmín Belén Quiroga, Julián Ansaldo, Luciano Lapenna, Santiago Bezchinsky, Raul Barriga Rubio, and Mailén García. Doing the feminist work in ai: Reflections from an ai project in latin america. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery.

[16] David Betancur Sánchez, Nuria Aldama García, Álvaro Barbero Jiménez, Marta Guerrero Nieto, Patricia Marsà Morales, Nicolás Serrano Salas, Carlos García Hernán, Pablo Haya Coll, Elena Montiel Ponsoda, and Pablo Calleja Ibáñez. Mel: Legal spanish language model, 2025.

[17] Víctor Mireles Chávez, Mariana Esther Martínez Sánchez, Javier Yankelevich Winocur, and Gerardo Sánchez Nateras. Buscando a los desaparecidos de la "guerra sucia": ontologías computacionales y la búsqueda de verdad. *Iberoforum. Revista de Ciencias Sociales*, 1(1):1–40, may 2021.

[18] Torsten Hiltmann, Martin Dröge, Nicole Dresselhaus, Till Grallert, Melanie Althage, Paul Bayer, Sophie Eckenstaler, Koray Mendi, Jascha Marijn Schmitz, Philipp Schneider, Wiebke Sczeponik, and Anica Skibba. Ner4all or context is all you need: Using llms for low-effort, high-performance ner on historical texts. a humanities informed approach, 2025.

[19] Bowen Zhang and Harold Soh. Extract, define, canonicalize: An llm-based framework for knowledge graph construction, 2024.

[20] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2025.

[21] Yi-Fan Lu, Xian-Ling Mao, Tian Lan, Heyan Huang, Chen Xu, and Xiaoyan Gao. Beyond exact match: Semantically reassessing event extraction by large language models, 2025.

[22] Equipo Angelus. Proyecto angelus: Primer conjunto de datos público, November 2022.

[23] Google DeepMind. Gemini Pro. `https://deepmind.google/models/gemini-pro/`, 2025. Consultado el 9 de junio de 2025.

[24] Gemma Team. Gemma 3 technical report, 2025.

[25] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.