Proyecto integrador:
Avance 1. Análisis exploratorio de datos

Misael López Sánchez A01796906

Proyecto Integrador MNA 2026
Equipo 18

Partner:
Dra. Anthony Rios

Dra. Grettel Barceló Alonso
Dr. Luis Eduardo Falcón Morales

25/Enero/2026

Table of Contents

Tabla de figuras

Lista de tablas

1. Introducción

La integración de Grandes Modelos de Lenguaje (LLMs) en el ámbito biomédico representa una de las fronteras más prometedoras de la inteligencia artificial contemporánea. Sin embargo, su adopción en entornos críticos se ve obstaculizada por el fenómeno de las "alucinaciones" y la falta de garantías sobre la veracidad factual de sus respuestas. Este Proyecto Capstone, desarrollado a través de una vinculación estratégica de investigación entre el **Tecnológico de Monterrey** y **The University of Texas at San Antonio (UTSA)**, aborda esta problemática mediante la exploración de técnicas avanzadas de interpretabilidad y control mecánico conocidas como *Activation Steering*.

El objetivo central de la investigación es analizar y manipular las activaciones internas de un modelo de lenguaje (Llama 3) para alinear sus respuestas con evidencia científica consolidada. Para lograr este nivel de rigor, el proyecto se fundamenta en el uso de datos de alta integridad provenientes de la ***National Institute of Health (NIH)*** de los Estados Unidos.

El presente reporte, correspondiente al **Avance 1: Análisis e Ingeniería de Datos**, documenta el proceso exhaustivo de construcción del *dataset* experimental. Se detalla el flujo de trabajo técnico diseñado para extraer conocimiento estructurado de la base de datos **SemMedDB**, transformarlo en consultas de lenguaje natural y establecer una línea base de evaluación de desempeño del modelo. Este proceso de validación de datos constituye el cimiento indispensable para los experimentos de intervención neuronal que se desarrollarán en las etapas subsecuentes del proyecto.

2. Antecedentes

2.1 Obtención de los datos

Para la elaboración de este proyecto de investigación. El Dr. Anthony y su equipo realizaron utilizar una base pública del *National Institutes of Health (NIH)*. A continuación, se describen los pasos realizados desde la extracción de la información hasta el procesamiento de los datos el cual responde a la necesidad de contar con información curada y consolidada para la aplicación de técnicas Avanzadas de Steering.

2.2 La Institución: National Institutes of Health (NIH)

La infraestructura de datos proviene de la **National Institues of Health (NIH)** de los Estados Unidos. El NIH representa la agencia principal del gobierno estadounidense responsable de la investigación biomédica y de salud pública. Reconocida mundialmente por su rigor científico, esta institución administra vastos volúmenes de literatura médica y datos estructurados que sirven como estándar de oro para la investigación clínica y traslacional. El uso de datos avalados por el NIH garantiza que la información base del proyecto posee la validez y la estandarización necesarias para experimentos de Inteligencia Artificial en contextos críticos.

2.3 El Repositorio: SemMedDB

Específicamente, este estudio utiliza la base de datos **SemMedDB (Semantic MEDLINE Database)**. SemMedDB es un repositorio de predicaciones semánticas extraídas de los títulos y resúmenes de citas biomédicas de PubMed (la base de datos de literatura médica más grande del mundo).

La relevancia de SemMedDB para este proyecto radica en su estructura: utiliza el sistema **SemRep** para procesar lenguaje natural y extraer tripletas semánticas en el formato *Sujeto-Predicado-Objeto* (por ejemplo: ***Fármaco A [TRATA] Enfermedad B***). Esta estructuración convierte texto no estructurado en un grafo de conocimiento manejable, ideal para analizar la capacidad de direccionamiento (steering) en modelos de lenguaje.

2.4 Protocolo de Acceso y Adquisición

Es imperativo destacar que el acceso a *SemMedDB* está restringido y controlado para asegurar el uso ético y académico de la información. La adquisición del dataset no se realizó mediante descarga pública directa. Se ejecutó un protocolo formal de solicitud de acceso dirigido a la administración del NIH y la Biblioteca Nacional de Medicina (NLM), el cual fue aprobado para fines de esta investigación. Este proceso de validación institucional añade una capa de gobernanza de datos al proyecto, asegurando la

trazabilidad. El proceso desde la solicitud de acceso a los datos hasta su aprobación, y final descarga de la DB tomo un tiempo de dos semanas en el mes de Octubre de 2025

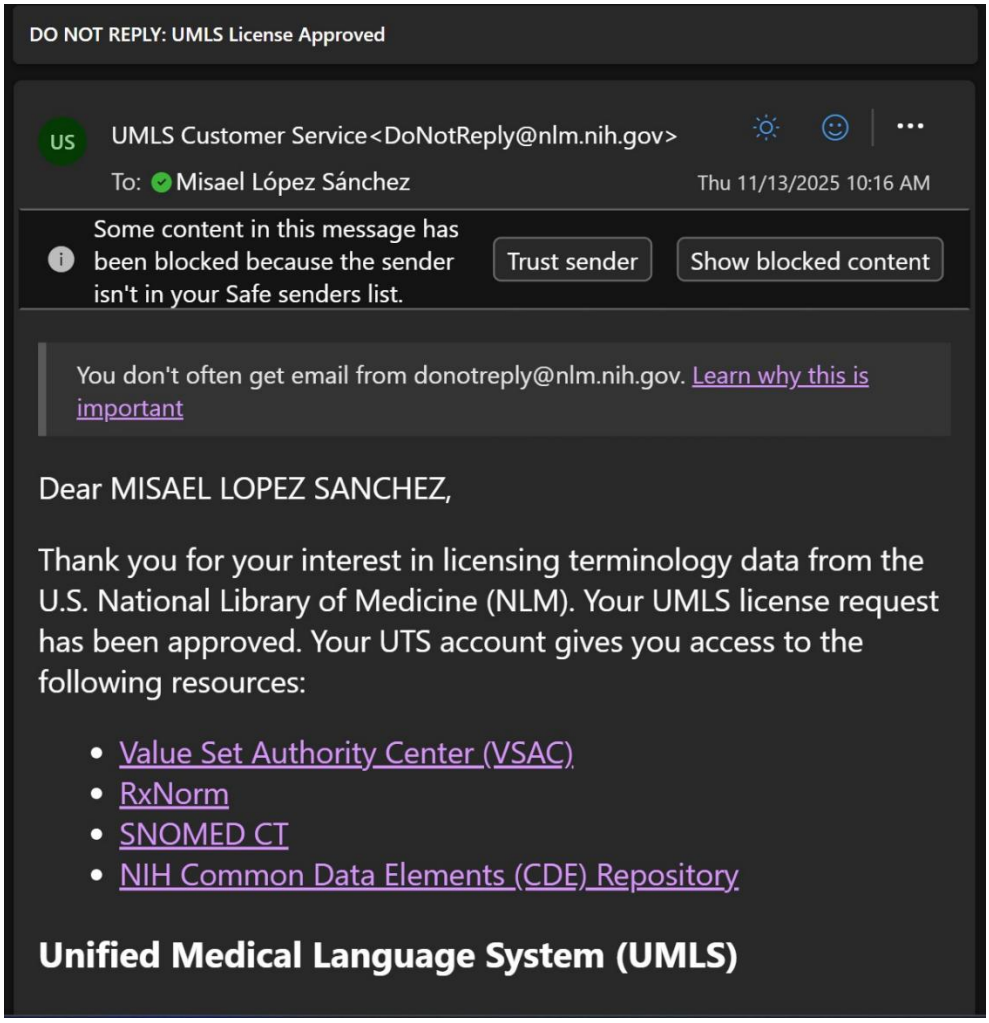


Ilustración 1 Evidencia de aprobación de solicitud UMLS

2.5 Esquema de Datos y Tablas Seleccionadas

Del conjunto total de SemMedDB, se seleccionaron y extrajeron dos tablas relacionales fundamentales para la construcción del dataset de entrenamiento y validación del proyecto: *CITATIONS* y *PREDICATIONS*. Las cuales son las tablas base sobre las cuales se construyó el proceso de Data Engineering para el proyecto. A continuación, se detalla el esquema y contexto de cada una:

A. Tabla CITATIONS (Contexto Bibliográfico)

Esta tabla actúa como la fuente de metadatos del documento original. Permite vincular cada hecho semántico con su origen en la literatura científica.

Campos de la tabla.

- **PMID (PubMed Unique Identifier):** Identificador único del artículo, que funge como llave primaria para relacionar la información.
- **ISSN:** Identificador de la publicación seriada (revista científica).
- **DP (Date of Publication):** Fecha de publicación, crucial para análisis temporales.
- **EDAT (Entrez Date):** Fecha en que la cita fue añadida a PubMed.
- **PY (Publication Year):** Año de publicación normalizado.

PMID	ISSN	DP	EDAT	PYEAR
19851774	1432-203X	2009 Dec	2010 01 21	2009

Ilustración 2 Estructura de la tabla CITATIONS

B. Tabla PREDICATIONS (Núcleo Semántico)

Esta es la tabla central para el análisis de *Activation Steering*, ya que contiene las relaciones lógicas extraídas del texto.

- **PREDICATION_ID:** Identificador único de la predicación extraída.
- **SENTENCE_ID:** Vincula la predicación a una oración específica dentro del abstract.
- **PMID:** Llave foránea que conecta con la tabla CITATIONS.
- **PREDICATE:** La relación semántica (el "verbo" o acción, ej. *TREATS, CAUSES, AFFECTS*).
- **SUBJECT_CUI / SUBJECT_NAME:** Identificador de Concepto Único (CUI) del Unified Medical Language System (UMLS) y el nombre textual del sujeto (ej. *Aspirin*).
- **OBJECT_CUI / OBJECT_NAME:** Identificador y nombre del objeto sobre el cual recae la acción (ej. *Headache*).

PREDICATION_ID	SENTENCE_ID	PMID	PREDICATE	SUBJECT_CUI	...	OBJECT_CUI	...	OBJECT_NOVELTY
1252467	3369924	16655556	AFFECTS	C1306232	...	C1326386	...	1

Ilustración 3 Estructura de la tabla PREDICATION

La integración de estas dos tablas permite no solo identificar *qué* se dice (Predications), sino *dónde y cuándo* se dijo (Citations), proporcionando un contexto robusto para el entrenamiento de los modelos.

2.6 Contexto del Esquema Relacional Complementario

Si bien el presente estudio se centra en las tablas CITATIONS y PREDICATIONS para la construcción del dataset de entrenamiento, es importante señalar que la arquitectura de **SemMedDB** se sustenta en un esquema relacional más amplio diseñado para garantizar la integridad y trazabilidad de la información semántica. A continuación, se describen brevemente las tablas auxiliares que completan el ecosistema de la base de datos:

- **SENTENCE:** Esta tabla almacena el texto crudo de las oraciones individuales extraídas de los resúmenes (abstracts). Su función principal es proporcionar la evidencia textual exacta de donde se derivó una predicación, sirviendo como "fuente de verdad" lingüística.
- **ENTITY:** Actúa como un puente de normalización. Mapea los fragmentos de texto específicos encontrados en una oración (ej. "ataque al corazón") con su Concepto Único Identificador (CUI) estandarizado en el UMLS (ej. *C0027051 - Myocardial Infarction*), detallando además su posición exacta y tipo semántico.
- **GENERIC_CONCEPT:** Contiene registros de conceptos que el sistema SemRep identificó como demasiado vagos o generales. Esta tabla es útil para filtrar ruido y asegurar que las relaciones extraídas tengan suficiente especificidad científica.
- **ACRONYMS:** Tabla de soporte destinada a la desambiguación, donde se registran las siglas y abreviaturas detectadas en el texto junto con sus expansiones o definiciones completas.

La omisión deliberada de estas tablas en el flujo de trabajo actual responde a la naturaleza del experimento de *Activation Steering*, el cual requiere relaciones lógicas consolidadas (tripletas) y metadatos temporales, prescindiendo de la reconstrucción textual oración por oración.

Como contexto general compartimos el mapa relacional de la base de datos completa.

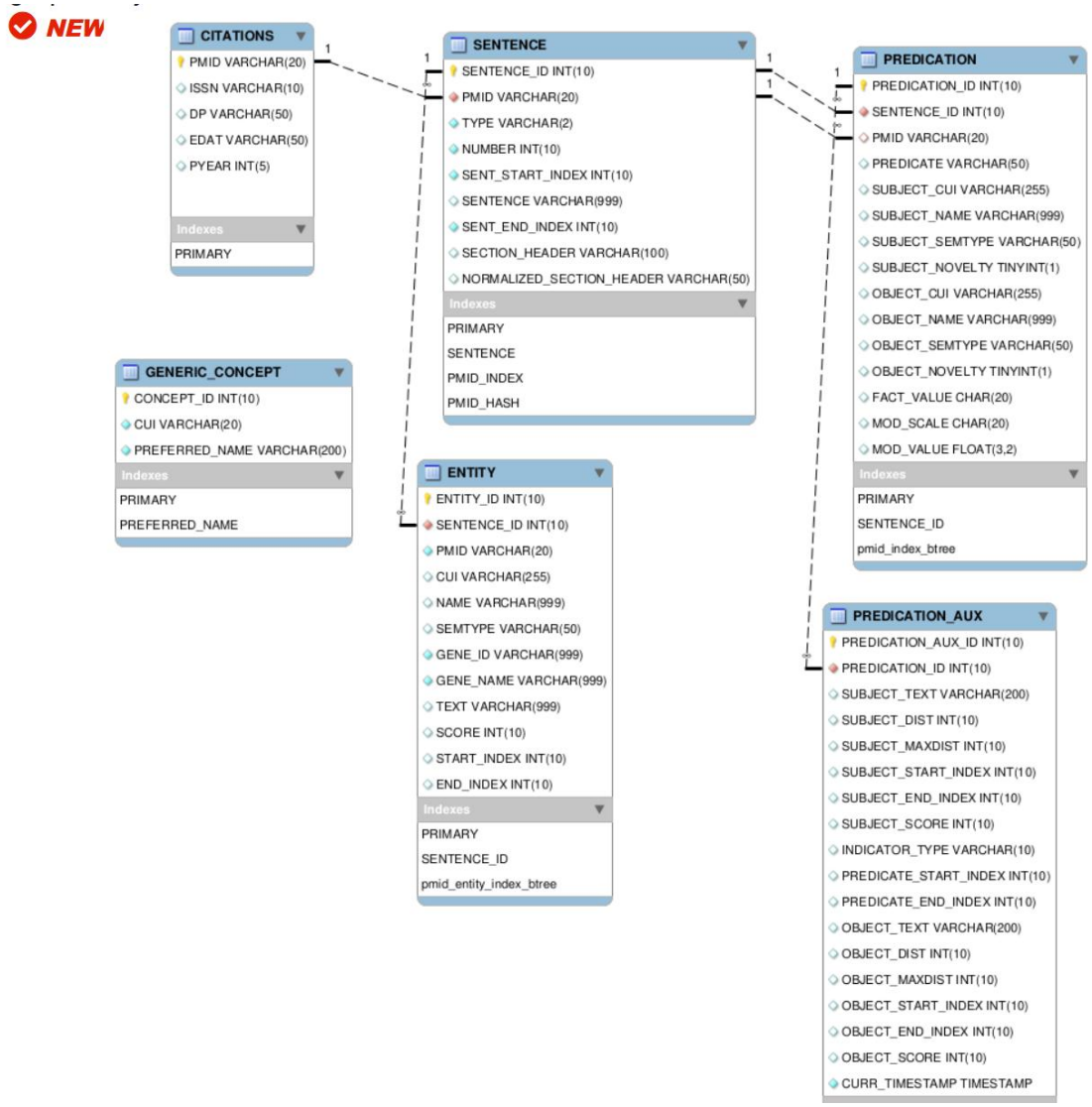


Ilustración 4 Diagrama de relación de la bas SemMedDB

Ahora se procederá a la explicación de la transformación de los datos en 3 diferentes fases.

3. Fase 1: Extracción y Construcción del Dataset de Tripletas Semánticas

3.1 Propósito del Módulo

El objetivo de esta primera etapa fue consolidar un dataset estructurado y manejable a partir de los datos crudos de SemMedDB (tablas CITATIONS y PREDICATIONS). Dado que el volumen total de SemMedDB abarca décadas y millones de registros, se diseñó un algoritmo de extracción selectiva para filtrar información relevante dentro de una ventana temporal específica (2010-2024) a petición de los doctores encargados del proyecto y asegurar un balance en la distribución de datos por año.

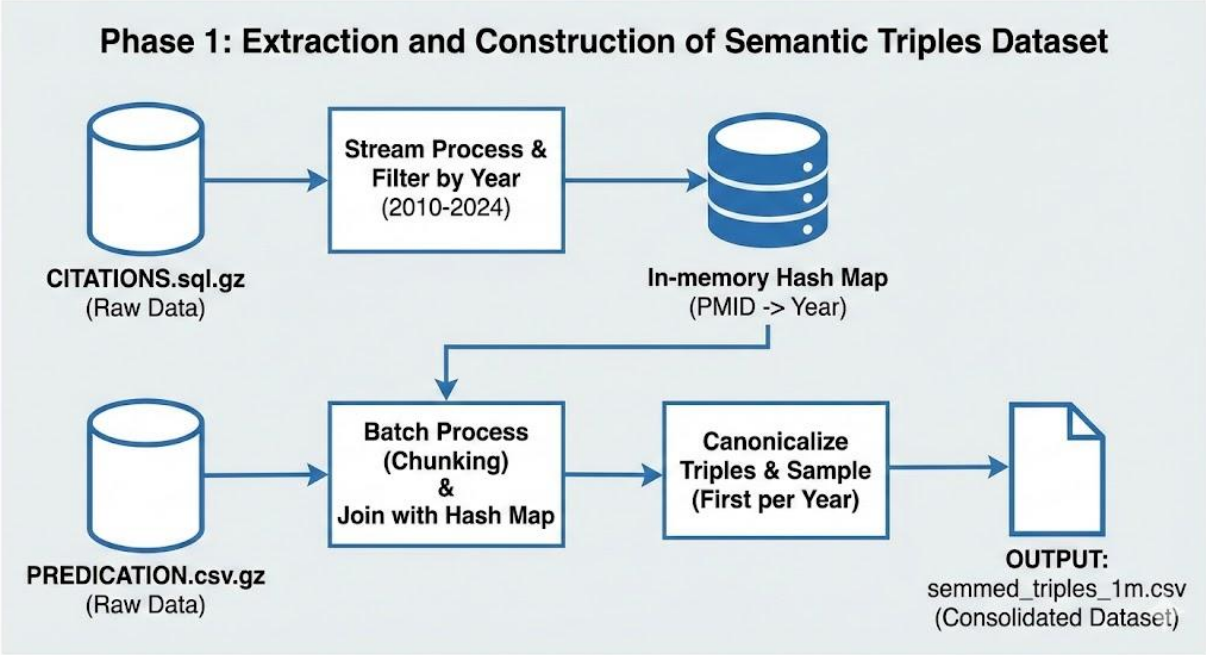


Ilustración 5 Flujo de extracción de datos y generación de triples

3.2 Metodología del Algoritmo de Procesamiento

Debido al tamaño masivo de los archivos fuente (que impide cargarlos completamente en la memoria RAM), se implementó un script en Python utilizando técnicas de **procesamiento por flujos (stream processing)** y **lectura por lotes (chunking)**. El flujo de ejecución consta de dos sub-procesos secuenciales:

A. Indexación Temporal (Procesamiento de CITATIONS) En primer lugar, se procesó el archivo de citas (*CITATIONS.sql.gz*) para construir un índice en memoria (Hash Map).

- **Lógica:** El algoritmo lee el archivo línea por línea utilizando expresiones regulares (*Regex*) para extraer únicamente dos datos clave: el identificador del artículo (PMID) y su año de publicación.
- **Filtrado:** Solo se indexaron aquellos artículos publicados dentro del rango objetivo (2010-2024). Esto actúa como un primer filtro, descartando literatura antigua que podría no reflejar el consenso médico actual o sesgar el modelo de lenguaje.
- **Resultado intermedio:** Un diccionario eficiente Map<PMID, Año> que permite consultar la fecha de un artículo en tiempo constante O(1).

B. Extracción y Canonicalización (Procesamiento de PREDICATIONS) Posteriormente, se procesó el archivo de predicaciones (PREDICATION.csv.gz) mediante la librería *Pandas*, leyendo el archivo en fragmentos (*chunks*) para no saturar la memoria.

- **Cruce de Datos (Join en Memoria):** Para cada predicación, el sistema verifica si su PMID existe en el índice temporal creado en el paso anterior. Si no existe (es decir, el artículo es muy antiguo o no está en el rango), el registro se descarta.
- **Canonicalización de Tripletas:** Un desafío común en SemMedDB es la redundancia (múltiples artículos mencionando el mismo hecho). Para mitigar esto, el algoritmo genera una "llave única" compuesta por: [Año, Predicado, Sujeto_CUI, Objeto_CUI].
- **Estrategia de Muestreo:** El script retiene únicamente la **primera aparición** de una tripleta única por año. Esto elimina duplicados exactos dentro del mismo periodo, asegurando diversidad en los datos.

3.3 Criterios de Selección y Filtrado

Para garantizar la calidad de los datos ingresados al pipeline de *Activation Steering*, se aplicaron las siguientes reglas de negocio automatizadas:

1. **Ventana de Observación:** Se restringió el análisis al periodo **2010 a 2024**, priorizando literatura biomédica contemporánea.
2. **Balanceo de Clases Temporal:** Se estableció un límite máximo de registros por año (definido en el parámetro PER_YEAR_TARGET, configurado a 70,000 registros en el script final) para evitar que los años más recientes, que suelen tener más publicaciones, dominen estadísticamente sobre los anteriores.
3. **Validación de Integridad:** Se descartaron registros que carecieran de identificadores semánticos (CUI) válidos para el sujeto o el objeto.

3.4 Esquema generado del mapeo de predicados

Los predicados que se hayan en la Base de datos son de diferente tipo. Aquí se mapean alguno de ellos para dar mayor claridad de los datos que contienen:

Tabla 1 Tabla de referencias de Predicados

Predicate	Meaning	Toy Example (Subject – Predicate – Object)
TREATS	A treats B	Aspirin TREATS Headache
PREVENTS	A prevents B	Vaccine PREVENTS Flu
CAUSES	A causes B	Virus CAUSES Fever
ASSOCIATED_WITH	A is correlated/related to B (generic)	Smoking ASSOCIATED_WITH Cancer
AUGMENTS	A increases effect of B	Caffeine AUGMENTS Alertness
DISRUPTS	A breaks/interferes with B	Mutation DISRUPTS Protein Function
STIMULATES	A increases activity of B	Exercise STIMULATES Heart Rate
INHIBITS	A reduces activity of B	Insulin INHIBITS Blood Sugar
AFFECTS	General influence	Stress AFFECTS Sleep
PREDISPOSES	A makes B more likely	Genetic Mutation PREDISPOSES Disease
COMPLICATES	A adds difficulty to B	Diabetes COMPLICATES Surgery
COEXISTS_WITH	A and B appear together	Fever COEXISTS_WITH Fatigue
CONVERTS_TO	A transforms to B	Glucose CONVERTS_TO Energy
PRODUCES	A makes B	Bacteria PRODUCES Toxin
USES	A uses B	Doctor USES Stethoscope
PROCESS_OF	A is a process of B	Digestion PROCESS_OF Nutrition
PART_OF	A is contained in B	Nucleus PART_OF Cell
LOCATION_OF	B located in A	Lungs LOCATION_OF Oxygen Exchange
ISA	A is a type of B (hypernym)	Penicillin ISA Antibiotic
MEASURES	A measures B	Thermometer MEASURES Temperature
DIAGNOSES	A diagnoses B	X-ray DIAGNOSES Fracture
ADMINISTERED_TO	A is given to B	Vaccine ADMINISTERED_TO Child
INDICATES	A suggests B	Pain INDICATES Injury
RESULT_OF	A happens because of B (inverse of CAUSES)	Fever RESULT_OF Infection

MANIFESTATION_OF	Symptom of disease	Cough MANIFESTATION_OF Cold
PRECEDES	A happens before B	Infection PRECEDES Fever
INTERACTS_WITH	A interacts with B	Drug A INTERACTS_WITH Drug B
OCCURS_IN	A happens in B	Inflammation OCCURS_IN Tissue
ASSOCIATED_WITH_INFER	Associated via inference (only in inference)	Gene X ASSOCIATED_WITH_INFER Cancer

3.5 Output Generado: semmed_triples.csv

El resultado de este proceso es un archivo CSV consolidado y ordenado, que sirve como la fuente de verdad para las siguientes etapas. Este archivo transforma las relaciones abstractas en registros legibles.

Esquema del Archivo Resultante:

- **first_year:** Año de publicación de la cita.
- **predicate:** La relación semántica (Ej. *TREATS, CAUSES*).
- **subject_cui / subject_name:** Identificador y nombre del concepto origen (Ej. *C0030193 / Penicillin*).
- **object_cui / object_name:** Identificador y nombre del concepto destino (Ej. *C0004093 / Bacterial Infection*).
- **first_pmid:** Referencia al artículo científico original que valida esta relación (trazabilidad).

Ejemplo de output del archivo resultante.

Tabla 2 Tabla de referencia del output fase1

Field	Value
subject_cui	10057
subject_name	ABCC5
subject_semtype	gngm
predicate	ADMINISTERED_TO
object_cui	C0553257
object_name	Renal Cell
object_semtype	cell
first_year	2010
first_pmid	19903828
citation_pmid	19903828
first_issn	*(vacío)*

first_sentence_id	19546579

4. Fase 2: Generación de Consultas en Lenguaje Natural (NLG)

4.1 Propósito del Módulo

Una vez consolidado el dataset de tripletas, la segunda fase del flujo de trabajo se centró en la **generación de estímulos textuales**. El objetivo de este módulo es transformar cada relación semántica abstracta (Sujeto Predicado Objeto) en una interrogante gramaticalmente coherente en idioma inglés. Estas preguntas son fundamentales para el proyecto, ya que servirán como *prompts* de entrada para calcular los vectores de *Activation Steering* en el modelo Llama 3, permitiendo analizar cómo la red neuronal representa la información biomédica interna.

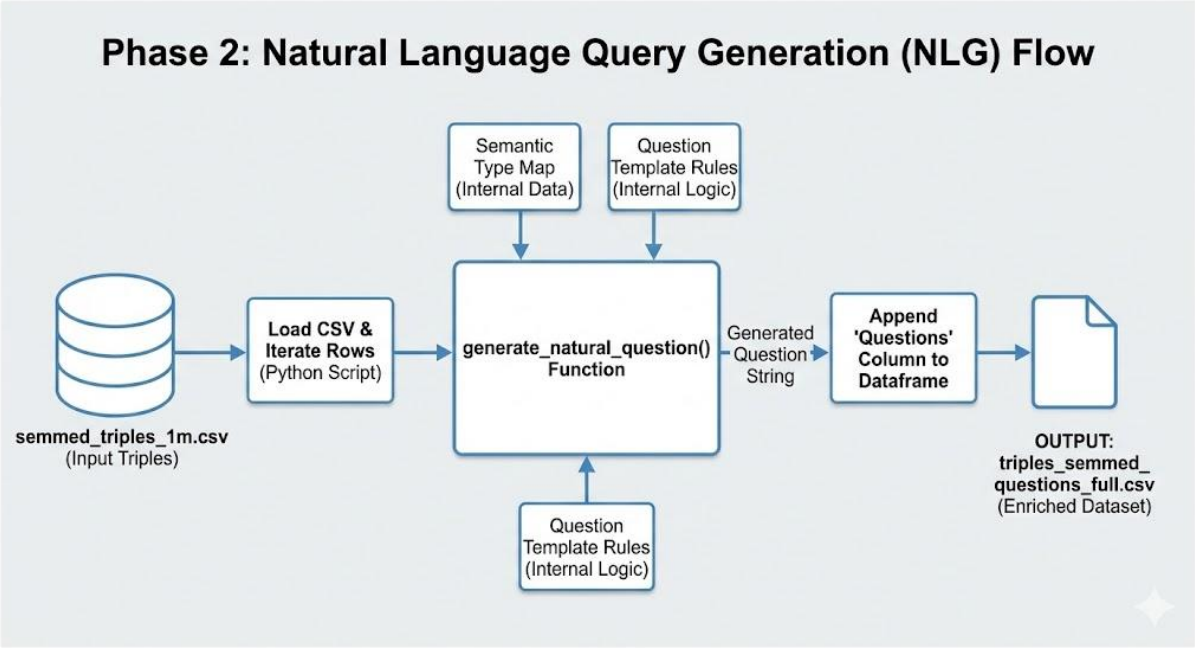


Ilustración 6 Flujo de generación de preguntas usando NLG

4.2 Metodología del Algoritmo de Generación

El script *generate_questions_semm.py* implementa un motor de generación de texto basado en reglas semánticas y plantillas contextuales. A diferencia de una traducción literal, este proceso requiere interpretar los metadatos médicos para construir oraciones con sentido clínico. El procesamiento se divide en tres pasos:

A. Decodificación de Tipos Semánticos (Semantic Mapping) La base de datos SemMedDB utiliza códigos abreviados de cuatro letras (ej. phsu, dsyn) para categorizar las entidades. Para que las preguntas sean naturales, el algoritmo integra un diccionario de mapeo (SEM_TYPE_MAP) que traduce estos códigos técnicos a terminología humana legible.

- ***Transformación:*** **phsu ->** "pharmacologic substance" (Sustancia farmacológica).
- ***Transformación:*** **humn ->** "patient group" (Grupo de pacientes).
- ***Impacto:*** Esto permite que el modelo entienda el contexto de la entidad, diferenciando si el objeto de la oración es un virus, un medicamento o un procedimiento quirúrgico.

B. Construcción Sintáctica de Preguntas El núcleo del script es la función *create_natural_question*. Esta rutina analiza la triada (Sujeto, Predicado, Tipo_Semántico_Objeto) y selecciona la estructura gramatical más adecuada para formular la pregunta.

- **Lógica de Plantillas Dinámicas:** Dependiendo del predicado (ej. *TREATS, CAUSES, ADMINISTERED_TO*), el algoritmo ajusta la preposición y el fraseo.
    - *Caso Específico:* Si el predicado es ADMINISTERED_TO, la pregunta se formula orientada al receptor: *"To which [Semantic Class] is [Drug Name] typically administered?"*.
    - *Caso General:* Para relaciones más abstractas, se utiliza una formulación explícita que contextualiza la consulta: *"In a biomedical context, what is the relationship between [Subject] and [Object Class] via [Predicate]?"*.

C. Reordenamiento y Consolidación Finalmente, el script asegura que la nueva información se integre correctamente al dataset sin perder la trazabilidad. Se implementa una lógica de reordenamiento de columnas para insertar el campo Questions inmediatamente después de los identificadores de la oración (first_sentence_id), facilitando la inspección visual y el procesamiento posterior.

4.3 Output Generado: triples_semmed_questions.csv

El producto de esta fase es un dataset enriquecido que mantiene los campos del paso anterior pero añade *Questions* que son respuestas generadas por el mismo LLM que utilizará en el procesamiento de Steering.

Tabla 3 Tabla de referencia ouput Fase 2

Field	Value
subject_cui	C0011900
subject_name	Diagnosis
subject_semtype	hlca
predicate	ADMINISTERED_TO
object_cui	C0030705
object_name	Patients
object_semtype	humn
first_year	2010
first_pmid	27755799
citation_pmid	27755799
first_issn	*(vacío)*
first_sentence_id	16904164
Questions	To which patients or groups is Diagnosis typically administered

5. Fase 3: Inferencia con LLM y Evaluación Automatizada de Respuestas

5.1 Propósito del Módulo

El objetivo de esta fase fue someter al modelo de lenguaje (Llama 3) a un interrogatorio sistemático utilizando las preguntas generadas en la fase anterior. El propósito no es entrenar al modelo, sino **evaluar su conocimiento intrínseco** sobre las relaciones biomédicas extraídas de SemMedDB. Esta evaluación permite clasificar cada tripleta en dos categorías: "Conocimiento Retenido" (el modelo responde correctamente) o "Alucinación/Desconocimiento". Esta distinción es vital para la futura extracción de vectores de *steering*, ya que permite diferenciar entre activar un conocimiento existente o inyectar uno nuevo.

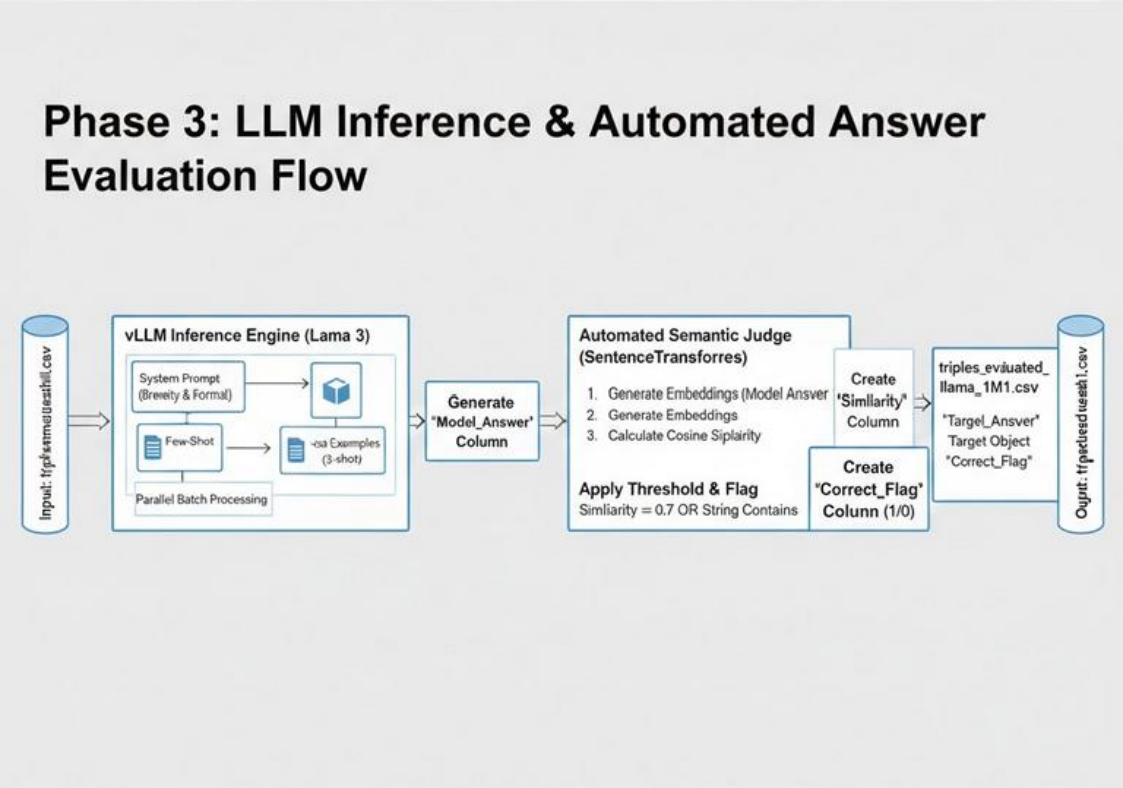


Ilustración 7 Flujo de la Etapa 3 y generación de respuestas

5.2 Metodología de Inferencia y Evaluación

El script ***llama_vllm_answer_evaluate.py*** implementa un pipeline de inferencia de alto rendimiento y un sistema de "juez automatizado". El proceso se estructura en tres componentes técnicos:

A. Configuración del Motor de Inferencia (vLLM) Para procesar el gran volumen de preguntas de manera eficiente, se utilizó la librería vLLM optimizada.

- **Estrategia de Prompting (Few-Shot Learning):** Se implementó una técnica de aprendizaje de *pocas oportunidades* (3-shot). Al modelo no solo se le presenta la pregunta, sino que se le proporcionan 3 ejemplos previos de par *Pregunta-Respuesta* (definidos en FEW_SHOT_DB).
- **Restricción del System Prompt:** Se inyectó una instrucción de sistema estricta para forzar la brevedad: *"Answer with ONLY the specific medical term... Do not use full sentences."* Esto es fundamental para evitar que el modelo divague y para facilitar la evaluación automatizada.

B. Evaluación de Similitud Semántica (Semantic Judge) Dado que el lenguaje natural es variable (ej. el modelo puede responder "Heart Attack" mientras que la base de datos dice "Myocardial Infarction"), una comparación exacta de texto (string matching) sería insuficiente. Por ello, se implementó un evaluador semántico utilizando SentenceTransformers (S-BERT).

- **Cálculo de Similitud:** Se generan *embeddings* vectoriales tanto de la respuesta del modelo (Model_Answer) como del objeto objetivo (Target_Object). Se calcula la **similitud coseno** entre ambos vectores.
- **Umbral de Decisión:** Se definió un umbral de corte (SIMILARITY_THRESHOLD = 0.7). Si la similitud supera este valor, o si existe una contención directa de cadenas, la respuesta se marca como correcta.

5.3 Resultados y Descripción del Dataset Final: triples_evaluated_llama.csv

El resultado es un dataset enriquecido que no solo contiene la pregunta y la respuesta esperada, sino el desempeño real del modelo. Este archivo es el insumo final para el análisis de *activation steering*.

Descripción de las Variables Generadas:

- **Target_Object:** Es el "Ground Truth" o la respuesta correcta derivada de la tripleta original de SemMedDB (el objeto de la relación).

- **Model_Answer:** La respuesta textual cruda generada por Llama 3. Aquí observamos la capacidad del modelo para recuperar información biomédica sin acceso a herramientas externas.
- **Similarity (Métrica Continua):** Un valor flotante entre 0 y 1 que cuantifica qué tan cerca estuvo la respuesta del modelo de la respuesta esperada.
    - *Ejemplo:* Una similitud de 1.0 indica una coincidencia semántica perfecta (ej. "Stem cells" vs "Cells").
    - *Ejemplo:* Una similitud baja (ej. 0.27) indica una alucinación o respuesta incorrecta.
- **Correct_Flag (Métrica Binaria):** Una variable categórica (0 o 1) derivada de la similitud.
    - **1 (Correcto):** Indica que el modelo posee el conocimiento paramétrico de esa relación específica.
    - **0 (Incorrecto):** Indica que el modelo desconoce la relación o alucinó una respuesta errónea.

Esta segmentación es crítica: para los experimentos de *Steering*, a menudo interesa analizar las activaciones internas precisamente en los casos donde Correct_Flag = 0 (para corregir el modelo) o Correct_Flag = 1 (para reforzar el comportamiento).

Muestra de los datos del dataset final:

Tabla 4 Tabla de referencia output Fase 3 negativo

Field	Value
subject_cui	10057
subject_name	ABCC5
subject_semtype	gngm
predicate	ADMINISTERED_TO
object_cui	C0553257
object_name	Renal Cell
object_semtype	cell
first_year	2010
first_pmid	19903828
citation_pmid	19903828
first_issn	*(vacío)*
first_sentence_id	19546579
Questions	To which cell types is ABCC5 typically administered?
Model_Answer	Cancer cells
Similarity	0.4461038410663605
Correct_Flag	0
Target_Object	Renal Cell

Tabla 5 Tabla de referencia output fase 3 positivo

Field	Value
subject_cui	255022
subject_name	CALHM1
subject_semtype	aapp
predicate	ADMINISTERED_TO
object_cui	C0007634
object_name	Cells
object_semtype	cell
first_year	2010
first_pmid	19944073
citation_pmid	19944073
first_issn	*(vacío)*
first_sentence_id	23608135
Questions	To which cell types is CALHM1 typically administered?
Model_Answer	Pancreatic beta cells
Similarity	1
Correct_Flag	1
Target_Object	Cells

6. Conclusiones Generales del Análisis y Procesamiento de Datos

El desarrollo de este primer entregable ha permitido establecer una infraestructura de datos robusta, validada y alineada con los estándares de investigación biomédica necesarios para la colaboración entre el Tecnológico de Monterrey y UTSA. A través de la implementación del pipeline de tres fases, se derivan las siguientes conclusiones críticas:

1. Integridad y Relevancia Temporal del Dataset La estrategia de extracción aplicada sobre la base de datos SemMedDB (NIH) ha garantizado que el modelo se entrene exclusivamente con información científica vigente. Al restringir la ventana de observación al periodo **2010-2024** y aplicar un muestreo balanceado por año, se ha mitigado el sesgo de obsolescencia (información médica desactualizada) y el sesgo de frecuencia temporal. Esto asegura que los vectores de *steering* resultantes reflejarán el consenso médico contemporáneo.

2. Eficacia de la Transformación Semántica (Structured-to-NL) El módulo de Generación de Lenguaje Natural (Fase 2) demostró que es posible traducir relaciones lógicas rígidas (tripletas) en consultas naturales complejas sin perder precisión técnica. El uso de mapeos semánticos (SEM_TYPE_MAP) y plantillas dinámicas ha permitido

generar *prompts* que son gramaticalmente correctos para el LLM, pero que conservan la rigurosidad ontológica del UMLS, cerrando la brecha entre las bases de datos SQL y los modelos de Inteligencia Artificial Generativa.

3. Validación mediante Juez Semántico (S-BERT) El análisis de inferencia (Fase 3) evidenció que las métricas de evaluación tradicionales (coincidencia exacta de texto) son insuficientes para el dominio biomédico, donde múltiples términos pueden referirse al mismo concepto (ej. *Heart Attack* vs. *Myocardial Infarction*). La implementación de un evaluador basado en *embeddings* (SentenceTransformers) y similitud coseno ha validado exitosamente la capacidad de distinguir entre un error del modelo y un sinónimo válido. Esto proporciona una métrica de Correct_Flag confiable, esencial para separar los casos de "conocimiento retenido" de las "alucinaciones".

4. Preparación para Activation Steering Finalmente, el dataset consolidado (triples_evaluated_llama.csv) cumple con todos los requisitos para la siguiente etapa de experimentación. La segmentación binaria de las respuestas (Correcto/Incorrecto) permitirá aislar las activaciones neuronales específicas asociadas a la veracidad, permitiendo entrenar vectores de intervención que puedan potencialmente "dirigir" al modelo hacia respuestas factuales y reducir la tasa de alucinación en contextos médicos críticos.

7. Bibliografía.

Referencias Bibliográficas

Fuentes de Datos y Organismos

- Kilicoglu, H., Shin, D., Fiszman, M., Helsen, G., & Rindflesch, T. C. (2012). SemMedDB: A semantic MEDLINE database with multiscale predications. *BMC Bioinformatics*, *13*(1), 315. https://doi.org/10.1186/1471-2105-13-315
- National Institutes of Health. (s.f.). *Turning Discovery Into Health*. Recuperado el 31 de enero de 2026, de https://www.nih.gov
- National Library of Medicine. (s.f.). *SemMedDB Database Information and Schema*. Lister Hill National Center for Biomedical Communications. Recuperado de https://lhncbc.nlm.nih.gov/ii/tools/SemRep_SemMedDB_SKR/dbinfo.html
- National Library of Medicine. (s.f.). *Unified Medical Language System (UMLS)*. U.S. Department of Health and Human Services, National Institutes of Health. https://www.nlm.nih.gov/research/umls/index.html

Modelos de Inteligencia Artificial y Algoritmos

- AI at Meta. (2024). *The Llama 3 Herd of Models*. arXiv. https://doi.org/10.48550/arXiv.2407.21783
- Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., ... & Stoica, I. (2023). Efficient Memory Management for Large Language Model Serving with

PagedAttention. *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*. https://arxiv.org/abs/2309.06180
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

Software y Librerías de Procesamiento

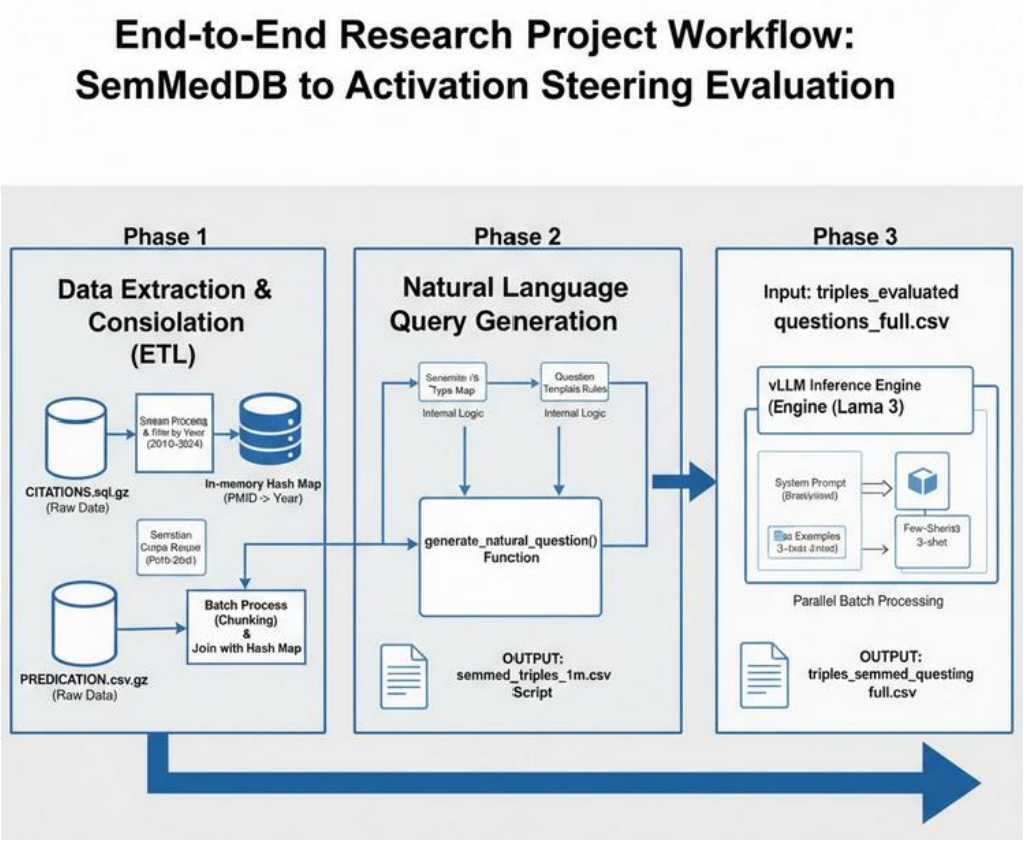
- Hugging Face. (s.f.). *Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX*. https://huggingface.co/docs/transformers
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56. https://pandas.pydata.org/

Anexos:

1.- Tabla con cada uno de los archivos principales sus insumos y archivos de salida.

Fase del Proyecto	Script de Procesamiento (Python)	Archivos de Entrada (Inputs)	Archivos de Salida (Outputs)
1. Ingeniería de Datos y Extracción	semmed_build_triples.py	• CITATIONS.sql.gz (Raw) • PREDICATION.csv.gz (Raw)	semmed_triples_1m.csv (Dataset base de tripletas limpias)
2. Generación de Lenguaje Natural (NLG)	generate_questions_semm.py	• semmed_triples_1m.csv	triples_semmed_questions.csv (Dataset enriquecido con columna 'Questions')
3. Inferencia y Evaluación de Conocimiento	llama_vllm_answer_evaluate.py	• triples_semmed_questions.csv • Modelo Llama 3 (vLLM)	*(Resultados de inferencia y evaluación, no especificados en detalle)*

2.- Flujo general de las 3 fases.



2.- Repositorio de github del proyecto.

Github Link:
https://github.com/MisaTecMNA/ProyectoIntegrador2026_Equipo18/tree/main