MEDICAL Q&A

Estefanía Laverde, Elmar Santofimio, Harvy Benítez

Contenido

01

Motivación

- Desarrollo de un sistema QA sobre dominio médico que comprenda la complejidad sobre razonamiento médico y que a su vez que provea respuestas bien definidas, confiables y entendibles
- Enfoque en la terminología médica especializada, razonamiento y representación legible de las respuestas sobre la búsqueda de información con base a las consultas de los usuarios
- Uso y experimentación de diferentes técnicas que permitan solventar los límites de los LLM en la actualidad como lo son las alucinaciones, límites sobre bases de datos privadas y enfoques generales que no tienen un desempeño relevante cuando se trata de temas especializados
- Dar respuestas complejas pero suficientemente comprensibles para un usuario que tenga un fuerte conocimiento especializado

02

Datasets

HEAD-QA

- Dataset constituido sobre 5.264 preguntas de respuesta múltiple en español (5 posibles opciones de respuesta) con diferentes especialidades de dominio médico (Medicina, Enfermeria, Psicología, Química, Farmacología y Biología) el cual pretende ser desafiante para usuarios con alto conocimiento especializado.

- Usado para evaluar las diferentes estrategias en la actual investigación.

- Este dataset ha sido evaluado en otras investigaciones arrojando líneas base para la precisión de 42%.

HEAD-QA

Accuracy por cada disciplina en el dataset con enfoque no supervisado en investigaciones previas

Model	Biology	Medicine	Nursing	Pharmacology	Psychology	Chemistry
Liu et al. (2020)	45.5	42.4	42.3	48.0	44.3	44.3
IR Baseline - Vilares and Gómez-Rodríguez (2019)	37.9	30.3	32.6	38.7	34.7	33.7

CoWeSe

- Es un dataset de alrededor de 4.5GB cerca de 750M tokens y cerca de 2 Millones de documentos médicos en texto plano y que abarca más de 3000 dominios en español. Es un dataset relevan te para datos biomédicos y del área de la salud que ha sido utilizado en previas investigaciones para la producción de embeddings.
- Se hará uso de este dataset continuamente en la investigación para:
  - Crear la base vectorizada que alimentara a su vez el RAG y el Agentic RAG.
  - Implementar Fine tuning para el modelo Decoder.

Nota: Por temas de recursos no se utilizara todo el dataset para la base de datos vectorizada.

Knowledge Graph (KG)

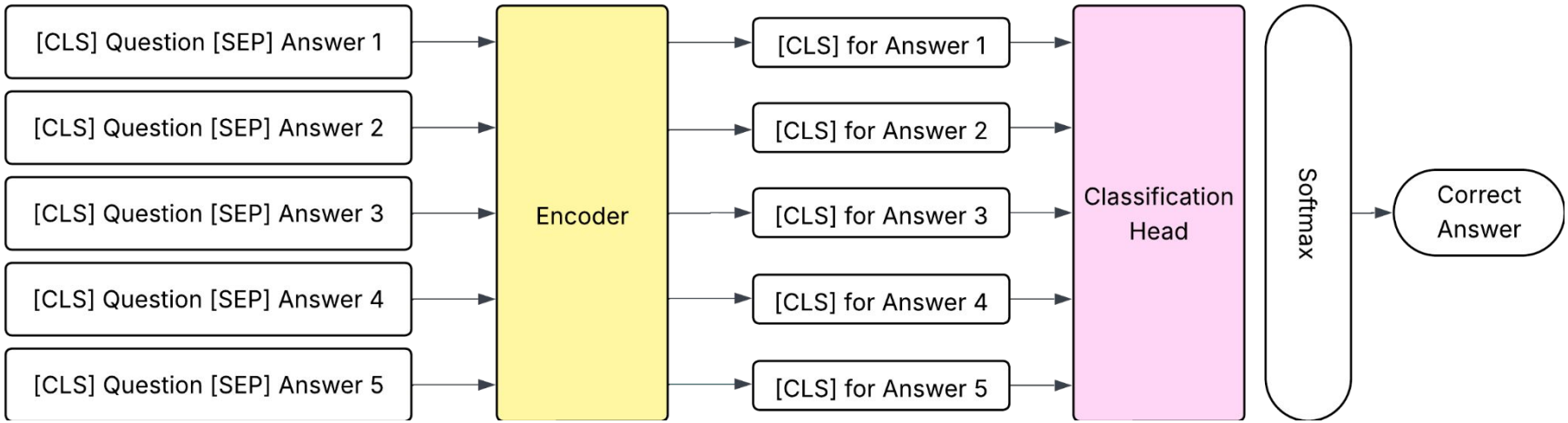
- Construido a partir de Wikidata, DBpedia y otras fuentes públicas.
  Contiene 1,140 nodos: 705 enfermedades, 160 síntomas, 184 tratamientos/procedimientos, 87 causas, 4 medicamentos.
  499 relaciones estructuradas, como HAS_SYMPTOM, TREATED_WITH, CAUSED_BY, USES_MEDICATION.
- Para crear y usar los nodos/relaciones se hace uso de Neo4j

Nota: Existen otras bases de datos verificadas y confiables como UMLS (EN) y Snomed CT (ES) sin embargo por temas de licencia internacional y que Colombia no está entre los miembros de la organización no fue posible utilizarlo.

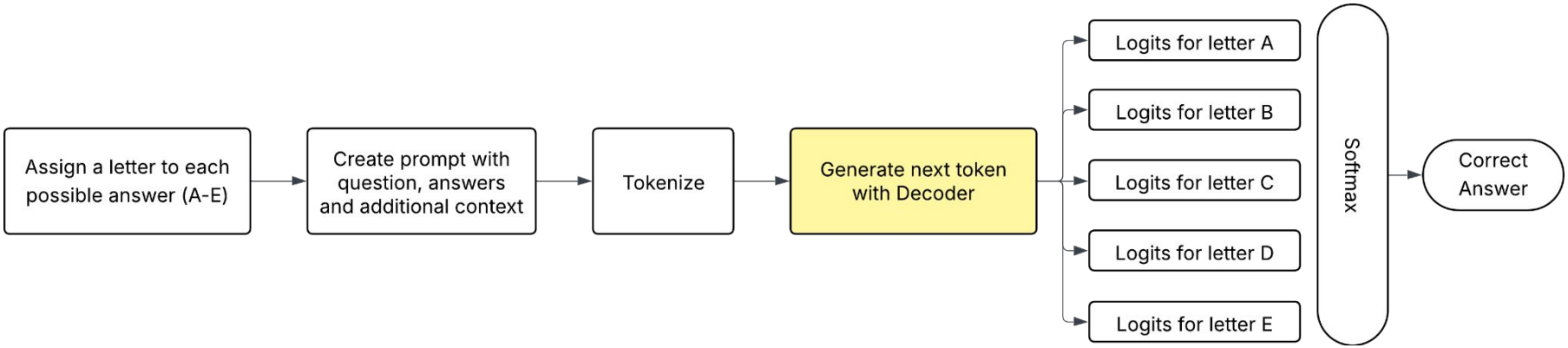
03

IMPLEMENTACIONES

Fine-tuning encoders en español en multiple-choice Q&A



Arquitecturas Decoder-Only para Selección de Respuestas Múltiples y Alternativo fine-tuning con LoRA



Arquitecturas Decoder-Only para Selección de Respuestas Múltiples y Alternativo fine-tuning con LoRA - Ejemplo de prompt

Eres un examinador experto en ámbitos biomédicos.
Debes seleccionar la opción correcta de las siguientes.
Responde únicamente con UNA letra (A, B, C, D o E) según corresponda.
Piensa paso a paso.

Pregunta: ¿Cuál de los siguientes agentes causa neumonía atípica?

Opciones:
A. Streptococcus pneumoniae
B. Mycoplasma pneumoniae
C. Haemophilus influenzae
D. Legionella pneumophila
E. Klebsiella pneumoniae

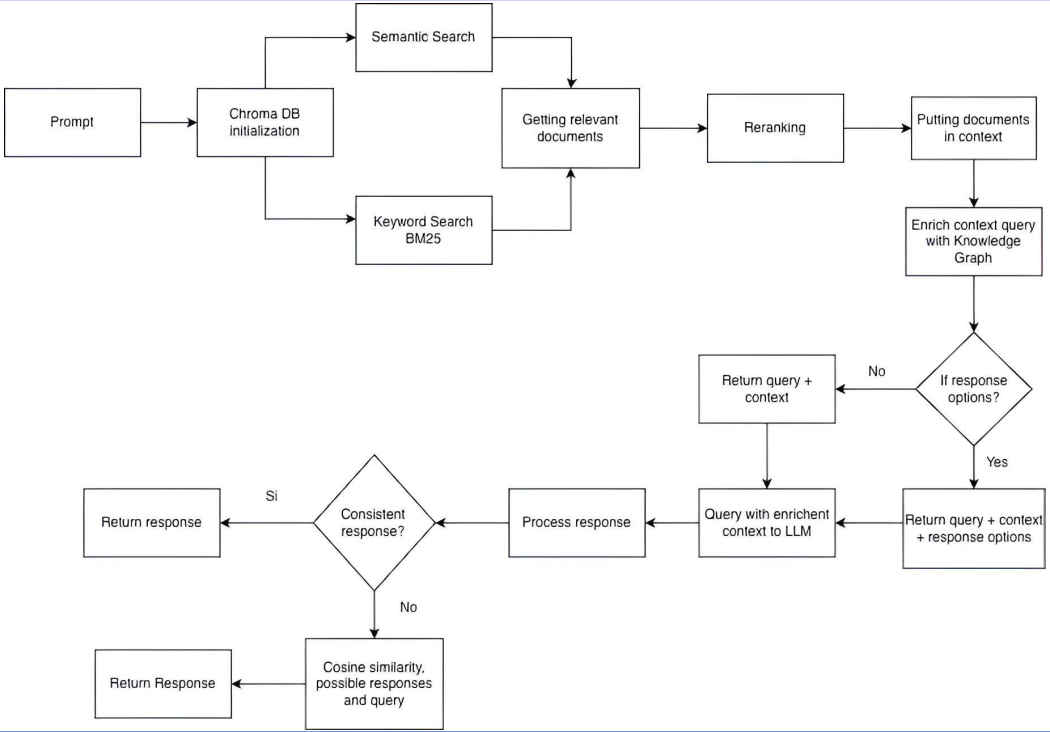
Contexto:...

Respuesta (solo la letra):

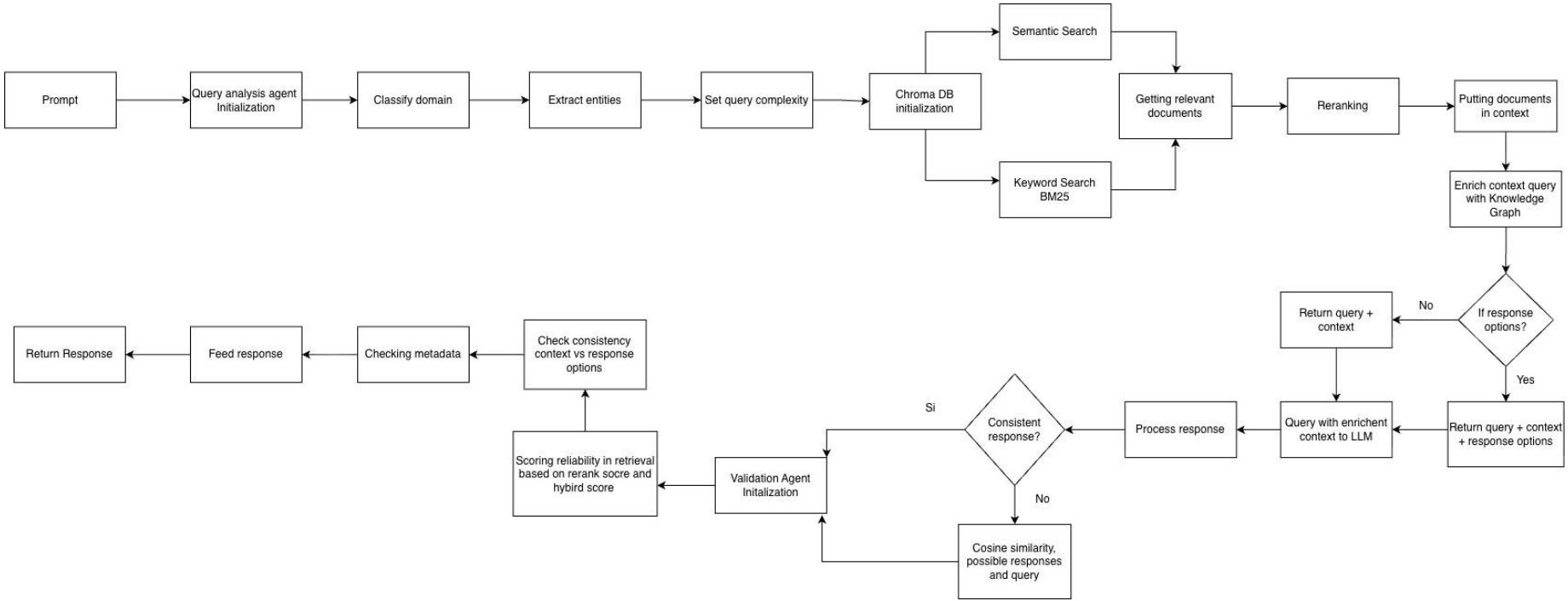
RAG INTEGRADO CON GRAFOS DE CONOCIMIENTO

- **Linea base:**
- La base de datos consta de 300.000 documentos sobre Chroma
  Motor de Búsqueda Híbrida combina dos técnicas complementarias:
  - ChromaDB para búsqueda vectorial semántica con un peso de 70% en la búsqueda
  - BM25 para búsqueda léxica por palabras clave con un peso de 30% en la búsqueda
- Grafo de Conocimiento Médico: Almacena relaciones entre enfermedades, síntomas, tratamientos y medicamentos gestionado en Neo4j.
- Adaptadores de LLM son tres los modelos comparados:
  - Llama 3.2 1B (eficiente, local)
  - Mistral 7B Instruct (balanceado)
  - Azure GPT-4o-mini (máximo rendimiento)

RAG INTEGRADO CON GRAFOS DE CONOCIMIENTO



AGENTIC RAG


Prompt
Query analysis agent Initialization
Classify domain
Extract entities
Set query complexity
Chroma DB initialization
Semantic Search
Keyword Search BM25
Getting relevant documents
Reranking
Putting documents in context
Enrich context query with Knowledge Graph
If response options?
No
Yes
Return query + context
Return query + context + response options
Query with enrichent context to LLM
Process response
Consistent response?
Si
No
Cosine similarity, possible responses and query
Validation Agent Initalization
Scoring reliability in retrieval based on rerank socre and hybird score
Check consistency context vs response options
Checking metadata
Feed response
Return Response

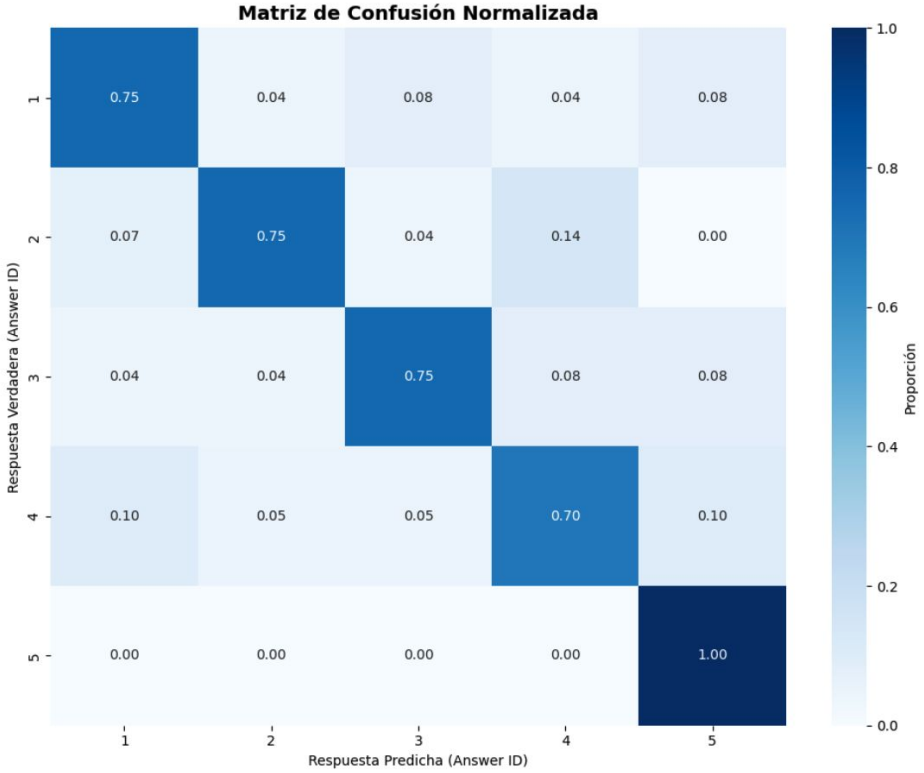
04

Resultados

Accuracy y F1-score

Método	Accuracy (%)	F1 (weighted) (%)
Fine-tune Bert	31.27	30.26
Decoder-Only en MC Q&A + contexto	54.05	53.73
RAG + KG (LLaMA 1B)	27.0	27.9
RAG + KG (LLaMA 7B)	45.0	45.9
RAG + KG (GPT-4)	77.0	77.7
Agentic RAG (Mistral 7B)	48.0	48.9
Agentic RAG (GPT-4)	75.0	75.7

Mejor matriz de confusión



05

CONCLUSIONES

- El sistema con GPT-4 + retrieval híbrido + knowledge graph alcanzó **77 % de precisión**, superando ampliamente los modelos más pequeños y los baselines previos.
- El enfoque A (fine-tuning de encoders) se comporta ligeramente mejor que una clasificación aleatoria, mientras que B (decodificadores generativos) logra **hasta 54 % de precisión**, lo que cual lo hace adecuado para entornos con recursos computacionales limitados.
- Los enfoques basados en recuperación (C y D) no solo mejoraron la precisión, sino que **ofrecieron evidencia para respaldar las respuestas**, lo cual es crucial en contextos médicos.
- El enfoque con agentes (C) generó mejoras marginales con modelos como Mistral 7B, gracias a pasos de razonamiento adicionales y verificación.
- Aunque su impacto directo en precisión fue moderado, contribuyó a respuestas más confiables y explicables, permitiendo afirmar hechos explícitamente (e.g., “la enfermedad X se asocia con Y”).
- La combinación de corpus no estructurado y grafos curados resulta prometedora para sistemas expertos de QA.

06

BIBLIOGRAFÍA

- D. Vilares and C. Gómez-Rodríguez (2019). HEAD-QA: A Healthcare Dataset for Complex Reasoning. In Proc. of ACL, 2019.
- C. P. Carrino et al. (2021). Spanish Biomedical Crawled Corpus: A Large, Diverse Dataset for Spanish Biomedical Language Models, 2021.
- X. Yang et al. (2024). Fine-tuning Medical Language Models for Enhanced Long-Contextual Understanding and Domain Expertise, 2024.
- J. Padilla Cuevas et al. (2024). MédicoBERT: A Medical Language Model for Spanish NLP Tasks with a QA Application Using Hyperparameter Optimization. Applied Sciences, vol. 14, no. 16, 2024.
- H. Cui et al. (2025). A Review on Knowledge Graphs for Healthcare: Resources, Applications, and Promises, 2025.
- D. Varshney et al. (2024). Knowledge-Grounded Medical Dialogue Generation Using Augmented Graphs. Scientific Reports, vol. 14, 2024.
- X. Dong et al. (2025). Talk Before You Retrieve: Agent-Led Discussions for Better RAG in Medical QA, 2025.
- G. Xiong et al. (2024). Benchmarking Retrieval-Augmented Generation for Medicine, 2024.
- R. S. Goodman et al. (2023). Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Network Open, vol. 6, no. 10, 2023.