Evaluation of LLMs with RAG for Factual Verification in the Biomedical Domain

université PARIS-SACLAY
INRAE MaIAGE

Lidan Zhang, Louise Deléger, Olivier Inizan, Arnaud Ferré
Paris-Saclay University, INRAE, MaIAGE, Jouy-en-Josas, France
Paris-Saclay University, CNRS, LISN, Gif-sur-Yvette, France

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
cnrs
LISN LABORATOIRE INTERDISCIPLINAIRE DES SCIENCES DU NUMÉRIQUE

I. Introduction

Large Language Models (LLM) have shown remarkable progress in **Natural Language Processing** (NLP), but their factual reliability in specialized conversational settings remains limited due to gaps in domain knowledge and susceptibility to **hallucinations**.

Retrieval-Augmented Generation[1] (RAG) is adressed to this limitation, which supplements the model's prompt with relevant information retrieved from an external Knowledge Base (KB).

Using a **biomedical** evaluation framework, we assess the **factual performance** of **10 conversational Large Language Models** on **two new Question-Answering** (QA) tasks with a study of the impact of RAG.

II. Materials

Dataset

ctd

*C*omparative *T*oxicogenomics *D*atabase

Chemical–Disease associations

represented as

Triplet		
Chemical	Predicate	Disease

C1 —is_related_to→ D1
C1 —is_related_to→ D2
C2 —is_related_to→ D3

turned in to natural language

Knowledge Base of Facts

C1 is related to D1.
C1 is related to D2.
C2 is related to D3.
...

LLMs

10 LLMs of varying sizes (7B–70B parameters)
All LLMs were queried via : **Ollama** or **Groq**
CoT: Chain **o**f **T**hought

Model	Parameters (B)	CoT	Open Weight
Gemma-7b	7	No	Yes
Mistral-8b	8	No	Yes
Qwen-7b	7	No	Yes
Llama3-8b	8	No	Yes
Gemma2-9b-it	9	No	Yes
Llama-4-maverick-17b-128e	17	No	Yes
Llama-4-scout-17b-16e	17	No	Yes
Mistral-saba-24b	24	No	No
Deepseek-r1-distill-llama-70b	70	Yes	Yes
Llama-3.3-70b-versatile	70	No	Yes

III. Prompting with RAG

Knowledge Base (KB)
Melatonin is related to Acromegaly.
Melatonin is related to Hypotension.
Melatonin is related to Seizures .
Amoxicillin is related to Diarrhea.
Lithium is related to Nausea.
...

Embedding Model → **KB Embeddings** —FAISS[2] Retriever→ **Retrieved Facts**

Query
Q1: *Is it true that Melatonin is related to Seizures?*
Q2: *What diseases are Melatonin related to?*

Embedding Model → **Query Embeddings** —FAISS Retreiver→ Retrieved Facts

Retrieved Facts —filtered by a threshold→ **Enhanced Prompt**

Query: user's question expressed in natural language, to be matched against the KB.
Embedding model: converts both the query and the knowledge base facts into vectors.

Document Retrieval

FAISS Retrieval: Effectively retrieves Top-K most semantically similar facts from the Knowledge Base using vector similarity (ANN search).
Top-K: Predefined as 1 for TFUQ and 30 for OEQ.

Threshold Filter: keeps only facts with cosine similarity ≥ 0.6, proved the most effective value in our experience.
Enhanced Prompt: combines the query with retrieved and filtered facts.
⇒ Providing the LLM with relevant context for more accurate answers.

Two QA settings, each 100 queries

OEQ (Open Ended Questions): enumerate all objects associated with a given Chemical and Disease;
TFUQ (True-False-Unknown Questions): ask the model to judge the veracity of a triplet with a ternary answer.

OEQ
Here are some known facts:
Melatonin is related to Acromegaly.
Melatonin is related to Hypotension.
Melatonin is related to Seizures.
List all diseases related to this chemical substance: Melatonin. Write one disease name per line, with numbering. Do not add any explanation or extra text.

→ LLM →

Expected response:
1. Acromegaly
2. Hypotension
3. Seizures

TFUQ
Here are some known facts:
Melatonin is related to Acromegaly.
Based on these: Is it true that Melatonin is related to Seizures? Answer only True or False. If you don't know, answer Unknown.

→ LLM →

Possible response:
True. / False. / Unknown.

Response Generation

These outputs are then compared to the expected responses from the KB to assess accuracy.

we expect answers to shift toward *True* since the relevant facts are injected.

IV. Prompting without RAG

Query
Q1: *Is it true that Melatonin is related to Seizures?*
Q2: *What diseases are Melatonin related to?*

→ **Prompt**

OEQ
List all diseases related to this chemical substance: Melatonin. Write one disease name per line, with numbering. Do not add any explanation or extra text.

→ LLM →

Expected response:
1. Acromegaly
2. Hypotension
3. Seizures

TFUQ
Is it true that Melatonin is related to Seizures? Answer only True or False. If you don't know, answer Unknown.

→ LLM →

Possible response:
True. / False. / Unknown.

even if the correct answer is always *True* in this setting, an *Unknown* answer does not compromise veracity in the absence of internal knowledge.

Baseline setting: the LLMs receives the prompt directly, Unlike RAG, no external facts are retrieved or injected into the prompt. The model relies solely on its internal knowledge for response generation.

V. Results

QA Setting	TFUQ (True/False/Unknown%)		OEQ (F1-score)	
Model	**noRAG**	**RAG**	**noRAG**	**RAG**
Gemma-7b	**81/0/19**	**100/0/0**	0.014	0.968
Mistral-8b	**12/0/88**	96/0/4	0.007	0.957
Qwen-7b	35/8/57	96/1/3	0.011	0.760
Llama3-8b	59/9/32	99/0/1	0.029	0.950
Gemma2-9b-it	**51/0/49**	**100/0/0**	0.031	0.991
Llama-4-maverick-17b-128e	68/5/27	**100/0/0**	0.087	0.986
Llama-4-scout-17b-16e	52/3/45	**100/0/0**	0.041	0.967
Mistral-saba-24b	**36/0/64**	99/0/1	0.031	0.971
Deepseek-r1-distill-llama-70b	28/17/55	99/0/1	0.047	0.989
Llama-3.3-70b-versatile	72/1/27	**100/0/0**	0.078	0.946

For TFUQ: performance was measure by the distribution of answers across {*True, False, Unknown*}. Without RAG, we hypothesized that hallucinations occur when the model answers *False*. Without RAG, performance varies across models, with no advantage for larger architectures over smaller ones. *Gemma-7b* and *Gemma2-9b-it*, perform factually well by assigning *Unknown* while still giving a relatively high proportion of *True* and no *False*. With RAG, 5 models correctly answer all questions while the others show a substantial increase in *True* responses (≥ 96).

For OEQ: performance was measured by F1-score (the harmonic mean of precision and recall in listing diseases related to each chemical). Scores are extremely low without RAG (F1 < 0.1) but rise sharply with RAG (F1 > 0.94) for all models.

VI. Conclusion

Our evaluation of querying a biomedical KB highlights the positive role of RAG:

Without it, all models struggled, while its integration led to a improvement in performance in both QA settings.
Gemma-7b and Gemma2-9b-it stood out, suggesting that even models of modest sizes can be competitive in terms of factual accuracy and be highly effective when paired with RAG.

This work presents preliminary single-hop experiments. Future work will extend to **multi-hop queries**, focusing on **Chemical-Gene-Disease** relations from the CTD.

C1 —interacts_with→ G1 —is _associated_with→ D1
C1 —is_related_to→ D1

VII.References

1. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In NeurIPS, 2020.
2. J. Johnson, M. Douze, and H. Jégou. Billion-scale Similarity Search with GPUs. In IEEE Transactions on Big Data, 2017.