Evaluating Factual Grounding Strategies in Large Language Models

Pablo Fernández, Anxo Pérez, and Javier Parapar

Information Retrieval Lab (IRLab), Faculty of Computer Science, Universidade da Coruña, 15071 A Coruña, Spain
Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: p.fernandezf@udc.es

DOI: *https://doi.org/10.17979/spu.23.c55*

Abstract:
Large Language Models (LLMs) often generate non-factual, or "hallucinated," content, limiting their reliability in knowledge-intensive tasks. This challenge is particularly critical in multi-hop question answering (MHQA), where models must integrate and reason over multiple pieces of evidence. In this paper, we present an empirical study of prompting strategies aimed at improving the factual grounding of LLMs. Using the Llama-3.1 (8B) model on the HotpotQA benchmark, we evaluate five prompting techniques along three main design points: shot count (zero-shot vs. few-shot), context integration (with vs. without supporting documents), and output constraints (free-form vs. structured responses requiring evidence). We assess both answer accuracy and the precision of supporting fact identification, allowing us to analyze correctness from evidential grounding. Our results reveal different trade-offs: while few-shot prompting improves reasoning consistency, the gains diminish without high-quality supporting context. Similarly, structured outputs reduce variance and improve factual alignment, but their benefits depend critically on how evidence is presented and constrained. Compared to prior studies that focus primarily on accuracy, our analysis highlights the importance of balancing answer quality with verifiable evidence. These findings provide actionable guidance for the design of prompts in multi-hop QA and inform broader efforts to mitigate hallucinations in LLMs across retrieval-augmented and reasoning-intensive applications.

Introduction

Question Answering (QA) has long been a central task in Natural Language Processing (NLP), evolving from early information retrieval–based systems that relied on keyword matching and structured resources Ferrucci et al. (2010); Voorhees and Tice (2000) to neural architectures capable of directly extracting answers from text Rajpurkar et al. (2016). In recent years, the emergence of Large Language Models (LLMs) has transformed this landscape, enabling generative models that can synthesize answers across diverse contexts Achiam et al. (2023); Bao et al. (2024); Chowdhery et al. (2023). Despite their remarkable capabilities, LLMs face a critical limitation: their tendency to produce non-factual or "hallucinated" content Ji et al. (2023). Unlike extractive QA systems, which point directly to source spans, generative models often obscure the provenance of their outputs, making verification and trustworthiness difficult. This challenge is particularly acute in knowledge-intensive tasks where factual reliability is essential Lewis et al. (2020); Petroni et al. (2019).

To address this, several datasets have been designed to evaluate not only answer correctness but also evidential grounding. Among them, multi-hop question answering (MHQA) has emerged as a particularly challenging setting. In MHQA, the answer is not located in a single passage but must be derived by combining information from two or more sources, often

through explicit reasoning steps such as bridging facts or making comparisons. For example, to answer "Which band was founded earlier?", the system must extract the founding year of each band from separate documents and then reason over these facts to reach the correct conclusion. Benchmarks such as HotpotQA Yang et al. (2018) and BeerQA Qi et al. (2021) are designed precisely to test these capabilities, requiring models to integrate disparate pieces of evidence rather than relying on shallow span extraction. These tasks simulate more realistic reasoning scenarios and expose the brittleness of LLMs when factual grounding is weak.

Prior work on mitigating hallucinations has explored retrieval-augmented generation (RAG) Lewis et al. (2020), instruction tuning Wei et al. (2022), and structured prompting Kojima et al. (2022). However, while retrieval methods emphasize evidence inclusion, less is known about how specific prompting strategies affect factual grounding in multi-hop QA. Most studies focus primarily on answer accuracy, with limited attention to the trade-off between correctness and evidential precision. In this work, we present a systematic evaluation of prompting strategies aimed at improving factual grounding in LLMs. Using the HotpotQA benchmark, we investigate five prompting approaches along three main points of design: shot count (zero-shot vs. few-shot), context integration (with vs. without supporting documents), and output constraints (free-form vs. structured formats requiring explicit evidence). To control for model variability, we fix the architecture to Llama-3.1 (8B) Grattafiori et al. (2024), which provides open weights and a manageable computational footprint, enabling reproducibility on widely available hardware. Our contributions are threefold:

- We provide a systematic comparison of prompt-based factual grounding strategies in multi-hop QA.
- We analyze the effects of answer accuracy from evidential precision, offering a more nuanced evaluation of grounding.
- We highlight practical insights into how context quality and output structuring shape the stability and reliability of LLM responses.

Related Work

Question Answering (QA) has been approached from two main paradigms: extractive and generative methods. Extractive models locate answers directly in the provided context, typically by predicting text spans, as in early neural approaches such as BiDAF Seo et al. (2017) or BERT-based readers Devlin et al. (2019); Gabín et al. (2021). By contrast, generative methods synthesize answers in natural language, potentially diverging from the surface form of the supporting text. While this generative flexibility improves fluency and adaptability, it also introduces the risk of hallucinated content Ji et al. (2023). The release of the HotpotQA dataset Yang et al. (2018) marked a turning point for evaluating systems that must perform reasoning across multiple documents. Unlike single-hop QA benchmarks such as SQuAD Rajpurkar et al. (2016), HotpotQA explicitly requires evidence integration through comparison or bridging facts. Several extensions and related datasets have since emerged, including BeerQA Qi et al. (2021), and 2WikiMultihopQA Welbl et al. (2018), each aiming to probe models' ability to handle increasingly complex reasoning chains. Leaderboards for HotpotQA track both answer accuracy and supporting fact identification, underscoring the dual importance of correctness and grounding.

The problem of hallucination has motivated multiple lines of research. Retrieval-augmented generation (RAG) approaches incorporate external evidence during inference, improving factuality in knowledge-intensive QA Lewis et al. (2020). Other efforts focus on prompting strategies, such as chain-of-thought reasoning Kojima et al. (2022), or structured answer formats Wei et al. (2022), which constrain model outputs to include evidential grounding. Instruction tuning and alignment techniques Ouyang et al. (2022) further refine models' tendencies to follow prompts faithfully. While prior work has extensively explored retrieval and fine-tuning

for factuality, relatively fewer studies have systematically compared prompting strategies for multi-hop QA. Most evaluations emphasize answer accuracy, whereas the trade-off between correctness and evidential precision remains less understood. Our work fills this gap by benchmarking a suite of prompting approaches for LLMs on HotpotQA, studying the effects of shot count, context integration, and output constraints on both answer quality and factual grounding.

Proposal

Multi-hop Question Answering (MHQA) represents one of the most demanding forms of QA, as it requires models to integrate evidence from multiple documents and perform reasoning steps such as bridging intermediate facts or comparing attributes across entities. This makes it an ideal benchmark for studying the factual reliability of LLMs, which are prone to hallucinations when reasoning chains grow longer or distractor information is present. In our work, we propose a systematic evaluation of prompting strategies for MHQA. Our experiments are designed to study three critical dimensions of prompt design: *i*) Shot count, whether models benefit from zero-shot vs. one-shot examples. *ii*) Context integration, whether grounding improves when supporting documents (and distractors) are explicitly provided. *iii*) Output structuring, whether constraining models to produce answers with evidential support reduces hallucinations and improves stability.

Dataset and Experimental Setting

For our experiments, we rely on the HotpotQA dataset Yang et al. (2018), which provides two evaluation settings: *fullwiki* and *distractor*. We adopt the *distractor* configuration, in which questions are paired with two gold paragraphs that contain the necessary evidence, along with eight automatically retrieved distractor paragraphs. The distractors are obtained using bigram TF–IDF retrieval over Wikipedia, with the question as query, and then mixed with the gold paragraphs (Yang et al., 2018, p.4).

We consider this setting particularly appropriate for benchmarking factual grounding, as the presence of distractors requires the model not only to locate correct evidence but also to ignore misleading content. This makes the task more challenging than closed-domain QA and more representative of real-world knowledge-intensive reasoning.

Dataset Statistics

We conduct all experiments on the development split of HotpotQA, which consists of 7,405 multi-hop questions, all labeled as "hard." Each entry contains the fields shown in Table 1, including the question, gold answer, supporting facts, and a context consisting of gold and distractor documents. Questions fall into two types: bridge questions, which require chaining intermediate facts, and comparison questions, which involve reasoning over attributes from multiple entities.

Methods

We evaluate five prompting strategies to study the impact of shot count, context integration, and output structuring on factual grounding. All experiments use the Llama-3.1 (8B) model Grattafiori et al. (2024), accessed via a local Ollama instance through the OpenAI Python API. For structured outputs, we define Pydantic schemas, which are converted into Grammar-Constrained Generation (GBNF) grammars to guide token sampling during generation Geng et al. (2023).

Table 1: Fields in the HotpotQA development split.

_id	Entry identifier
question	Question text
context	Context available as `[["Document title", [list of lines]], ...]`
answer	Answer text
supporting_facts	List of supporting facts in the form `[["Document title", line_index], ...]`
type	Question type ("bridge" or "comparison")
level	Question difficulty ("hard" for all questions in the used set)

Raw prompting. As a baseline, we use direct prompting with the `transformers` library, without Ollama or output structuring. The model is asked only to provide a natural-language answer.

Answer-only Ollama. In this configuration, the LLM is prompted through Ollama with a Pydantic schema containing a single field, `answer`, described as *"strictly the answer to the question."*. This tests whether minimal structure stabilizes output formatting without constraining evidence.

Complete Ollama (zero-shot). Here, the schema includes both an `answer` field and a `supporting_facts` field. The model must not only provide the answer but also identify the relevant evidence lines from the context. To guide formatting, the schema is shown directly in the prompt, but no example entries are provided.

Complete Ollama (one-shot). This configuration extends the previous setup with a one-shot example. The example illustrates both the answer and the supporting facts, with the latter underlined in the provided context. Table 2 shows the illustrative entry we designed for this experiment. The inclusion of a worked example is intended to help the model internalize the process of selecting evidence rather than only mimicking output format.

Results

Evaluation Metrics

Following the official HotpotQA evaluation protocol Yang et al. (2018), we report performance using Exact Match (EM) and F_1 score. EM measures the proportion of predictions that exactly match the gold answer, while F_1 computes token-level overlap between prediction and gold. To compute F_1, the model outputs are normalized by:

- Lowercasing all answers.
- Stripping punctuation and stopwords (e.g., a, an, the).
- Handling special cases for yes/no questions and "noanswer" predictions, which are returned directly without further token-level comparison.

We evaluate both answer accuracy (EM/F1) and, where applicable, supporting fact identification (SupFac EM/F1), as well as a joint metric that requires both the answer and supporting facts to be correct. All results are reported on the HotpotQA distractor development split, which includes 7,405 hard multi-hop questions.

Table 2: Example entry used in the one-shot configuration. Supporting facts are underlined in the context. A third distractor document is also included.

Question	Which association was created earlier, the Association of Tennis Professionals or the Women's Tennis Association?
Context	Association of Tennis Professionals: - The Association of Tennis Professionals (ATP) is the governing body of the men's professional tennis circuits (...) - It was formed in September 1972 by (...) Women's Tennis Association: - The Women's Tennis Association (WTA) is the principal organizing body of women's professional tennis. - The association governs the WTA Tour (...) - The WTA's corporate headquarters are in St. Petersburg, (...) - The Women's Tennis Association was founded in June 1973 (...) International Tennis Federation: - The International Tennis Federation (ITF) is the governing body of world tennis, wheelchair tennis, and beach tennis. - It was founded in 1913 as the International Lawn Tennis Federation by twelve national tennis associations.
Answer	Association of Tennis Professionals

Results

Table 3 summarizes the results across the four prompting strategies under two settings: with and without access to context paragraphs (gold + distractors). Several clear trends emerge from the results:

Context is essential. Across all methods, the inclusion of supporting documents and distractors significantly improves answer accuracy. For example, Answer-only improves from EM 0.1410 (no context) to 0.2505 (with context), and F1 from 0.2078 to 0.3500. Without context, the model struggles to answer meaningfully, often defaulting to irrelevant or generic responses.

Structured outputs improve stability. Comparing Answer-only to Complete (zero-shot), we see a modest gain when supporting facts are required: EM rises from 0.2505 to 0.2966, and F1 from 0.3500 to 0.4096. Although supporting fact EM/F1 remains low, the requirement of evidence appears to encourage more consistent reasoning.

Few-shot prompting provides further gains. The Complete one-shot method yields the best performance overall, reaching EM 0.3372 and F1 0.4571 for answers, with notable improvements in supporting fact identification (SupFac F1 = 0.1622 vs. 0.0908 for zero-shot). This suggests that even a single illustrative example helps the model learn how to align answers with evidence.

Joint accuracy remains low. While one-shot prompting improves evidence grounding, the joint EM/F1 scores remain under 0.01 and 0.10 respectively. This highlights the difficulty of

Table 3: Experiment results on the HotpotQA distractor development split. Answer = answer accuracy, SupFac = supporting fact prediction, Joint = both correct.

	Method	Context	Answer		SupFac		Joint	
			EM	F1	EM	F1	EM	F1
	Raw prompting	No	0.0013	0.0482				
		Yes	0.0017	0.0128	—	—	—	—
Structured	**Answer-only**	No	0.1410	0.2078				
		Yes	0.2505	0.3500	—	—	—	—
	Complete zero-shot	No	0.1446	0.2152				
		Yes	0.2966	0.4096	0.0084	0.0908	0.0024	0.0436
	Complete one-shot	No	0.1446	0.2152				
		Yes	**0.3372**	**0.4571**	**0.0170**	**0.1622**	**0.0089**	**0.0941**

requiring both the correct answer and the correct supporting facts simultaneously — a stricter and more realistic measure of factual grounding.

Overall, our experiments show that contextual grounding and output structuring both matter, but their benefits are maximized when combined with even minimal few-shot supervision. However, the low supporting fact and joint scores underscore that reliable evidence attribution in MHQA remains a major challenge for current LLMs.

Conclusions and Future Work

In this paper, we investigated the role of prompting strategies in improving the factual grounding of Large Language Models for multi-hop question answering (MHQA). Using the HotpotQA benchmark in its distractor setting and the Llama-3.1 (8B) model, we systematically compared four configurations that varied along three axes of design: shot count, context integration, and output structuring. Our findings reveal several key insights. First, providing context is indispensable: models without access to supporting documents fail to produce reliable answers. Second, structured outputs encourage stability and partial evidence alignment, though supporting fact identification remains challenging. Third, few-shot prompting yields the largest improvements, particularly in joint answer-and-evidence metrics, demonstrating that even minimal supervision can guide models toward better grounding.

Looking forward, there are several promising directions for future work. One avenue is to extend our study to larger and more diverse LLMs, including instruction-tuned and retrieval-augmented models, to examine how grounding strategies scale. Another is to explore richer supervision signals, such as chain-of-thought rationales or multi-modal evidence, to improve both reasoning transparency and factual alignment. Finally, we plan to investigate adaptive prompting frameworks that dynamically adjust structure and examples based on question type, aiming to reduce hallucinations while preserving model flexibility.

Bibliography

J. Achiam, S. Adler, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

E. Bao, A. Pérez, and J. Parapar. Adapting large language models for underrepresented languages. In *Proceedings XoveTIC 2024: Impulsando el talento científico*, pages 25–32, 2024.

A. Chowdhery, S. Narang, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

J. Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL 2019*, pages 4171–4186. ACL, June 2019.

D. Ferrucci, E. Brown, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

J. Gabín, A. Pérez, and J. Parapar. Multiple-choice question answering models for automatic depression severity estimation. *Engineering Proceedings*, 7(1):23, 2021.

S. Geng, M. Josifoski, M. Peyrard, and R. West. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the EMNLP 2023*, pages 10932–10952, Singapore, Dec. 2023. ACL.

A. Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

L. Ouyang et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

F. Petroni et al. Language models as knowledge bases? In *Proceedings of the EMNLP-IJCNLP 2019*, pages 2463–2473, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

P. Qi, H. Lee, O. T. Sido, and C. D. Manning. Answering open-domain questions of varying reasoning steps from text. In *Proceedings of the EMNLP 2021*, 2021.

P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. SQuAD: 100,000+ questions for machine comprehension of text. In J. Su, K. Duh, and X. Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, Nov. 2016. Association for Computational Linguistics. URL *https://aclanthology.org/D16-1264/*.

M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*, 2017.

E. M. Voorhees and D. M. Tice. Building a question answering test collection. In *Proceedings of the ACM SIGIR 2000*, pages 200–207, 2000.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

J. Welbl, P. Stenetorp, and S. Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the EMNLP 2018*, 2018.