Comparative Analysis of RAG vs. Fine-Tuning for Unanswerable Question Detection

Victoria Do and **Alejandro Franza** and **Eric Tsang**
{victoriavdo,afranza,ectsang}@berkeley.edu

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras ullamcorper elementum metus, eget dignissim felis posuere vitae. Nullam bibendum tellus sed bibendum gravida. Fusce posuere laoreet tortor, at facilisis magna imperdiet vel. Donec finibus, ipsum a finibus ornare, leo risus dictum felis, sit amet porta diam diam sit amet lectus. Aenean aliquam condimentum velit. Quisque sodales volutpat ligula, non gravida metus tincidunt in. Sed fermentum eros ac libero gravida, eu hendrerit enim porttitor. Fusce vehicula dignissim vehicula. Aenean tincidunt posuere lobortis. Nam sodales nunc id ipsum ullamcorper elementum. Nam molestie velit quis elit blandit placerat.

1 Introduction

Question answering (QA) represents a primary use of large language models (LLMs), especially in domains that require accurate information. However, QA systems face a fundamental challenge: they often hallucinate answers to unanswerable questions instead of abstaining (Ji et al., 2023).

This challenge becomes critical in high-stakes domains. Healthcare systems for clinical decision support must never fabricate medical advice, as hallucinations can have life-threatening consequences (Pal et al., 2023). Legal research systems that invent non-existence cases have resulted in attorney sanctions (Dahl et al., 2024). These domains particularly need robust unanswerable question detection because their knowledge bases have clear boundaries - medical guidelines cannot address every rare disease combination; legal databases cannot cover every novel situation.

Two paradigms address hallucination through different knowledge integration approaches. Retrieval-Augmented Generation (RAG) retrieves relevant passages before generation, providing external, non-parametric knowledge access (Lewis et al., 2020). Fine-tuning encodes knowledge directly into model parameters through training on domain-specific data, representing internal, parametric integration (Roberts et al., 2020).

Organizations deploying QA systems must choose between these paradigms yet lack empirical guidance on which better handles unanswerable questions. We address: *Do models more reliably abstain when knowledge is provided externally (RAG) or encoded internally (fine-tuning)? How does each paradigm balance answering correctly while recognizing knowledge boundaries?*

Gap in existing work. While prior work compares RAG and fine-tuning on standard QA metrics (Soudani et al., 2024), this focuses on accuracy when answers exist, not on abstention behavior. Hallucination detection work addresses fabrications after generation (Sadat et al., 2023; Farquhar et al., 2024), rather than comparing prevention through different knowledge integration approaches. SQuAD 2.0 studies concentrate on architectural improvements to extractive models (Rajpurkar et al., 2018), not paradigm comparisons for generative LLMs.

Our contribution. We present a systematic comparison of RAG versus fine-tuning focused on unanswerable question detection. Using SQuAD 2.0, we evaluate three systems: (1) zero-shot baseline with no retrieval or fine-tuning, (2) RAG system with dense retrieval from the SQuAD 2.0 corpus, and (3) Llama fine-tuned on SQuAD 2.0 using QLoRA. Our experimental design isolates the effect of knowledge integration paradigm on abstention behavior while maintaining comparable model capacity and training data. Beyond standard metrics (EM, F1), we measure answer grounding using BERTScore (Zhang et al., 2020) and Semantic Textual Similarity (STS) (Cer et al., 2017) to assess whether answers derive from provided/retrieved context versus parametric memory.

Results preview. Our evaluation reveals distinct tradeoffs between RAG and fine-tuning for unanswerable question detection...

2 Background

SQuAD datasets. Question answering systems have evolved significantly with large-scale benchmarks. Rajpurkar et al. (2016) introduced SQuAD 1.0 with 100,000+ answerable questions from Wikipedia. However, SQuAD 1.0 had a critical limitation: every question was guaranteed to be answerable from the given context, meaning systems never needed to recognize when they lacked sufficient information. To address this, Rajpurkar et al. (2018) released SQuAD 2.0, adding 50,000 unanswerable questions. This established evaluating both answer accuracy and the ability to recognize knowledge boundaries - the capability we use to assess how different knowledge integration methods handle unanswerable questions.

LLM hallucination. The tendency of LLMs to hallucinate—generate plausible yet unfactual content—is a fundamental challenge in deployment (Ji et al., 2023). While post-generation detection methods exist (Farquhar et al., 2024; Sadat et al., 2023), our work examines prevention through different knowledge integration paradigms—specifically comparing how RAG versus fine-tuning affect hallucination rates when questions are unanswerable.

Retrieval-Augmented Generation. Lewis et al. (2020) introduced RAG, which retrieves relevant documents using dense passage retrieval, then conditions generation on query and retrieved passages. This provides non-parametric knowledge access—information stored externally and provided at inference rather than encoded in weights. Shuster et al. (2021) found RAG reduces fabricated information in dialogue, suggesting external knowledge helps models distinguish accessible versus lacking information.

Fine-tuning. Fine-tuning represents an alternative paradigm where domain knowledge is encoded directly into model parameters through continued training on task-specific data. Roberts et al. (2020) demonstrated fine-tuning effectively injects factual knowledge into parameters for closed-book QA. This parametric approach internalizes knowledge within the model weights themselves.

PEFT. Parameter-efficient methods like LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2023) made this practical for large models by fine-tuning only small matrices while freezing pretrained weights. These enable training on domain data including unanswerable examples, potentially teaching boundary recognition. We adopt QLoRA in our fine-tuning experiments to ensure computational feasibility while maintaining model quality.

RAG-FT comparisons. Soudani et al. (2024) compared paradigms across knowledge-intensive tasks, finding strengths depend on task structure and that RAG excels with less popular knowledge while fine-tuning benefits from abundant training data. However, their evaluation emphasized answerable question performance using exact match and F1 scores, not abstention on unanswerable questions.

Our positioning. We address a gap in existing comparisons: systematic evaluation of how RAG versus fine-tuning handle unanswerable question detection in generative LLMs. While prior comparisons (Soudani et al., 2024) focus on accuracy when answers exist, we examine abstention behavior using SQuAD 2.0's unanswerable questions. Our controlled experimental setup compares three systems—zero-shot baseline, RAG with dense retrieval, and QLoRA fine-tuning—using the same base model and training data to isolate the effect of the knowledge integration paradigm. Beyond standard metrics (EM, F1), we use BERTScore (Zhang et al., 2020) and STS (Cer et al., 2017) to measure whether generated answers are grounded in provided/retrieved context versus hallucinated from parametric memory, directly quantifying the distinction between external and internal knowledge access.

3 Methods

Nunc dolor eros, lobortis egestas risus quis, facilisis ornare mi. Morbi a lacinia sapien. Etiam risus purus, lobortis vitae purus at, commodo condimentum tortor. Nulla diam metus, egestas vitae interdum ut, pretium sit amet nisl. Nam eget malesuada dui. Cras ultrices, felis aliquam ultricies iaculis, massa dolor semper turpis, quis placerat erat tortor et velit. Vestibulum posuere aliquam neque ut dignissim. Quisque mauris justo, euismod at convallis eget, aliquet eu lacus. Sed ut luctus neque. Praesent vel mauris sit amet odio ornare commodo. Maecenas lacinia iaculis magna, id porttitor quam tincidunt vitae. Vivamus et vehicula quam. Mauris malesuada euismod turpis, ut pellentesque tortor condimentum porttitor. Maecenas viverra lorem libero, in tristique ex iaculis non. Duis tempus lobortis lorem sed auctor.

4 Results and Discussion

Donec ut posuere quam. Nam eget dapibus erat, gravida viverra felis. Fusce eleifend urna sed risus suscipit, in vestibulum elit luctus. Vivamus ut leo convallis, interdum magna et, accumsan odio. Aenean vulputate purus at est sollicitudin, nec egestas justo iaculis. Fusce convallis sit amet nisi ac pharetra. Aenean maximus aliquam nibh eu ornare. Duis sit amet dictum magna, id malesuada urna. Curabitur cursus gravida odio. Donec eleifend iaculis vestibulum. Aliquam suscipit, lectus id malesuada ultricies, lorem ipsum malesuada nisi, sed venenatis magna metus non ante. Nulla a porttitor mauris, et pretium justo. Nulla dignissim mi non pharetra lacinia. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Interdum et malesuada fames ac ante ipsum primis in faucibus.

5 Conclusion

Cras cursus condimentum semper. Donec rhoncus ante ac mi aliquam rutrum. Nullam pulvinar felis vitae tellus porttitor, non vehicula ante consectetur. Fusce accumsan tortor at nunc finibus, non malesuada felis cursus. Nulla malesuada mattis mollis. Donec finibus consequat tincidunt. Quisque metus risus, blandit dictum mollis ut, molestie eget ex. Suspendisse mauris erat, ultrices vitae euismod ac, varius tempor nulla. Vivamus pellentesque nisl eu venenatis aliquam. Donec tincidunt ut lectus nec interdum. Nam eu libero luctus, dapibus odio ac, gravida quam. Quisque blandit quis velit vel consequat.

6 Authors' Contributions

Victoria Do contributed with the data preprocessing, exploratory data analysis (EDA), evaluation functions, and baseline model/results. Victoria was the primary author for the Introduction and Background sections. Victoria was also the primary author for the Data Description and Baseline Model Implementation subsections of the Methods section.

Alejandro Fernandez contributed…

Eric Tsang contributed…

References

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334, Singapore. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez, and Zhe Feng. 2023. DelucionQA: Detecting hallucinations in domain-specific question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 822–835, Singapore. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2024. Fine tuning vs. retrieval augmented generation for less popular knowledge. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2024, pages 12–22, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A First Appendix Title

This is appendix A content.

B Second Appendix Title

This is appendix B content.