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

Anonymous ACL submission

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 that becomes particularly significant in high-stakes applications: they often hallucinate answers to unanswerable questions instead of abstaining (Kalai et al., 2025).

Healthcare and legal domains are particularly critical. QA systems for consumer health queries, clinical decision support, and drug interaction checking must never fabricate information, as hallucinated medical advice can have life-threatening consequences (Pal et al., 2023). Similarly, systems for legal research, compliance checking, and contract analysis can cause penalties, liability, or rights violations through fabricated answers. Prominent scandals have occurred where AI systems invented non-existent legal cases, resulting in attorney sanctions (Dahl et al., 2024). These domains particularly need robust unanswerable question detection because knowledge bases have clear boundaries yet cannot cover every scenario. Medical guidelines cannot address every rare disease combination; legal databases cannot cover every novel situation. When questions fall outside a system's knowledge scope, it must recognize this rather than fabricate answers.

Two prominent paradigms address the hallucination challenge through different approaches to knowledge integration. Retrieval-Augmented Generation (RAG) retrieves relevant passages from a knowledge base before generation, representing an external, non-parametric approach to grounding responses (Lewis et al., 2020; Shuster et al., 2021). Fine-tuning directly encodes knowledge into model parameters through training on domain-specific question-answer pairs, including unanswerable examples, representing an internal, parametric approach (Roberts et al., 2020). While other mitigation techniques exist - including prompt engineering, chain-of-thought reasoning (Wei et al., 2022), and human-in-the-loop verification (Ouyang et al., 2022) - RAG and fine-tuning represent the two fundamental paradigms for integrating domain knowledge into QA systems.

Organizations deploying QA systems must choose between RAG and fine-tuning for knowledge integration, yet lack empirical guidance on which approach better handles unanswerable questions. Specifically, we address: *Do models more reliably abstain from answering when knowledge is provided externally through retrieval (RAG) or encoded internally through fine-tuning? How does each paradigm balance answering answerable questions correctly while recognizing when questions fall outside the knowledge scope?* These questions are critical for high-stakes deployments where incorrect answers carry significant consequences.

While prior work has compared RAG and fine-tuning on standard QA metrics like exact match and F1 scores, these comparisons focus primarily on accuracy when answers exist. Little empirical research examines how these paradigms differ in their ability to detect unanswerable questions and appropriately abstain (Soudani et al., 2024; Balaguer et al., 2024). Furthermore, existing hallucination detection work focuses on identifying fabricated content after generation, rather than comparing prevention strategies through different knowledge integration approaches (Sadat et al., 2023; Farquhar et al., 2024). Finally, studies on SQuAD 2.0 - a dataset explicitly designed to test abstention behavior - concentrate on architectural improvements to extractive QA models rather than comparing fundamental knowledge integration paradigms (Rajpurkar et al., 2018). Our work fills this gap through systematic evaluation of RAG versus fine-tuning specifically on unanswerable question detection.

We present a systematic comparison of RAG and fine-tuning paradigms specifically focused on unanswerable question detection using SQuAD 2.0. Beyond standard metrics, we introduce an answer attribution score that measures whether generated answers are grounded in provided/retrieved context versus hallucinated from parametric memory. This metric directly captures the key distinction between external and internal knowledge integration. We implement three systems using Llama models: (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 controlled experimental design isolates the effect of knowledge integration paradigm on abstention behavior while maintaining comparable model capacity and training data.

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

2 Background

Question answering systems have evolved significantly with the introduction of large-scale benchmarks. Rajpurkar et al. (2016) introduced the Stanford Question Answering Dataset (SQuAD 1.0), a reading comprehension benchmark containing over 100,000 questions where all answers could be extracted from provided Wikipedia passages. 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, augmenting the dataset with over 50,000 unanswerable questions. This established the paradigm of evaluating not just answer accuracy, but also the crucial ability to recognize knowledge boundaries—a capability we leverage to evaluate how different knowledge integration paradigms handle unanswerable questions.

Early approaches to unanswerable question detection in SQuAD 2.0 focused on architectural modifications to extractive models. These methods typically added answer verification components, threshold-based mechanisms, or modified BERT-based architectures to output “no answer” predictions when appropriate. However, this line of work concentrated on improving specific model architectures for extractive QA rather than examining fundamental differences in how knowledge is integrated into systems—the core question we address for modern generative LLMs.

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 Author's Contributions

Fusce pretium magna mauris. Cras mi elit, venenatis id turpis quis, vehicula ornare turpis. Vivamus molestie aliquet efficitur. Vestibulum ac massa justo. Sed vulputate, sem eu sollicitudin volutpat, erat metus tempor ante, in hendrerit ante diam a massa. Proin in fermentum libero. Donec nec quam a nulla tincidunt bibendum. Proin vitae lobortis odio. Vestibulum ut justo est. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Suspendisse vitae tincidunt nunc. Integer a urna sed massa pulvinar bibendum vitae ultricies est. Ut lacinia dolor purus, eget accumsan mauris tempor posuere. Morbi blandit ipsum elementum, fringilla ligula ac, viverra nibh.

References

Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O Nunes, Rafael Padilha, and 1 others. 2024. Rag vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint arXiv:2401.08406*.

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.

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

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *ArXiv*, abs/2509.04664.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. 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, page 12–22, New York, NY, USA. Association for Computing Machinery.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

A First Appendix Title

This is appendix A content.

B Second Appendix Title

This is appendix B content.

C Third Appendix Title

This is appendix C content.