

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

Question answering systems must recognize when questions are unanswerable to avoid hallucinating responses, yet empirical guidance on knowledge integration paradigms for abstention remains limited. We present a comparison of Retrieval-Augmented Generation (RAG) versus parameter-efficient fine-tuning (QLoRA) for unanswerable question detection, evaluating how external versus internal knowledge integration affects abstention behavior in generative LLMs. Using SQuAD 2.0 and Llama-3.2-3B-Instruct, our results reveal a critical instability in low-resource fine-tuning: While QLoRA peaked in unanswerable detection (74.53% EM), this came at the cost of functional performance, as it effectively stopped answering valid questions (2.99% EM). In contrast, the RAG system provided a robust alternative, improving unanswerable detection to 59.61% (vs. 52.01% baseline) while retaining functional answering capabilities (22.22% EM). Beyond standard metrics, we employ BERTScore and Semantic Textual Similarity to assess answer grounding. Our findings suggest that externalizing knowledge boundaries through retrieval yields more graceful degradation than parametric encoding, where models learn abstention as a dominant strategy, motivating future hybrid approaches.

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 adversely affect patient health and care quality (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 for QA over low-frequency entities (Soudani et al., 2024), their evaluation focuses on accuracy when answers exist in the knowledge base, not on abstention when questions are unanswerable. 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. We examine how the different knowledge integration paradigms affect abstention behavior, controlling for 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 context versus parametric memory.

Results preview. Our evaluation reveals critical behavioral differences between RAG and fine-tuning for unanswerable question detection. While both approaches improve abstention over zero-shot baseline of 52.01%, QLoRA fine-tuning learned abstention as a dominant strategy: it achieved the highest unanswerable detection score (74.53%) only by sacrificing nearly all utility on answerable questions (dropping to 2.99%). In contrast, RAG provided a more robust balance (59.61% unanswerable, 22.22% answerable). These results indicate that fine-tuning leads to overfitting on abstention. These findings motivate future work on hybrid approaches that combine retrieval’s external grounding with fine-tuning’s abstention behavior. Code is available at <https://github.com/viando8/datasci266-nlp-final-project/tree/main>.

2 Background

SQuAD datasets. QA 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. We leverage these unanswerable questions to evaluate how different knowledge integration methods handle abstention behavior.

LLM hallucination. The tendency of LLMs to hallucinate 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 unanswer-

able.

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 (Detrmers et al., 2023) made fine-tuning practical for large models by updating only small matrices while freezing pre-trained weights. These methods 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 these paradigms for QA over less popular factual knowledge, evaluating twelve language models on Wikipedia-based datasets using substring matching accuracy. They found RAG substantially outperforms fine-tuning across all popularity levels, with the largest gains for least popular entities. However, all questions in their evaluation were answerable from their data—they did not examine abstention behavior when questions are unanswerable.

Our positioning. We address a gap in existing comparisons: how RAG versus fine-tuning handle unanswerable question detection in generative LLMs. While prior comparisons (Soudani et al., 2024) evaluated QA performance when answers exist in the knowledge base, we examine abstention behavior using SQuAD 2.0’s unanswerable questions. Our 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

Table 1: SQuAD 2.0 dataset composition by answerability.

| Split | Type | Count |
|------------|--------------|----------------|
| Training | Answerable | 86,821 (66.6%) |
| | Unanswerable | 43,498 (33.4%) |
| | <i>Total</i> | <i>130,319</i> |
| Validation | Answerable | 5,928 (49.9%) |
| | Unanswerable | 5,945 (50.1%) |
| | <i>Total</i> | <i>11,873</i> |

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

3.1 Dataset Description

We evaluate all systems on SQuAD 2.0 (Rajpurkar et al., 2018), which includes 50,000 unanswerable questions alongside 100,000+ answerable ones. This dataset requires systems to determine both when they can answer and when they should abstain.

The dataset consists of question-answer pairs derived from Wikipedia articles across 477 distinct topics. Each example includes a context paragraph, a question, and either an extracted answer span or an empty array indicating the question is unanswerable. We use the official training split for model training and fine-tuning, and reserve the official validation split as our held-out test set.

Data Characteristics. Table 1 shows the dataset composition by answerability. The training set contains 130,319 examples with 66.6% answerable questions. In contrast, the validation set is evenly balanced with 50.1% unanswerable questions across 11,873 examples—a distribution that challenges models to recognize knowledge boundaries without relying on dataset bias.

Table 2 presents token count statistics for questions, contexts, and answers. Questions are consistently brief, with a median length of 12 tokens across both splits. Context paragraphs are substantially longer, with median lengths of 146 tokens (training) and 149 tokens (validation). Answer

Table 2: SQuAD 2.0 token count statistics using Llama-3.2-3B tokenizer. Mean values are shown with median in parentheses.

| Component | Split | Mean (Med.) | Range |
|-----------|-------|--------------|--------|
| Question | Train | 12.32 (12) | 1–61 |
| | Val | 12.37 (12) | 4–37 |
| Context | Train | 158.61 (146) | 26–907 |
| | Val | 166.90 (149) | 30–780 |
| Answer | Train | 4.69 (3) | 1–75 |
| | Val | 4.40 (3) | 1–42 |

spans for answerable questions are short, with a median of 3 tokens, though the distribution has a long tail extending to 75 tokens in the training set.

QLoRA subsets. Due to computational constraints, we train the QLoRA model on a randomly sampled subset of 8,000 training examples (6.1% of full data, seed=42) and evaluate on a 2,000-example subset of the validation split. This reduced scale may limit the model’s exposure to the full diversity of question types, a limitation we address in future work.

3.2 Baseline Model

Our baseline system answers questions from the SQuAD 2.0 validation dataset without retrieval or fine-tuning, using Llama-3.2-3B-Instruct (Grattafiori et al., 2024) in a zero-shot configuration. Each validation example is formatted with three components: (1) a task instruction defining expected behavior, (2) the context paragraph from SQuAD 2.0, and (3) the question.

Instruction Design. We evaluate three instruction variants on a stratified sample of 2,000 training examples to identify the optimal prompt formulation. Based on BERTScore F1 performance, we select the instruction variant designed to be more "explicit", emphasizing extractive behavior and providing clear abstention guidelines (see Appendix A for full methodology and selected instruction).

Decoding Strategy. The model must either extract an answer span from the context or output “unanswerable” when the context provides insufficient information. We employ low-temperature, top-p-constrained decoding (temperature=0.1, top_p=0.9, do_sample=True) to enforce extractive behavior, setting max_new_tokens=35 based on the distribution of answer token counts in the training dataset.

Output Parsing. Generated outputs are parsed by extracting text following the “Answer:” delim-

iter; responses containing keywords such as “unanswerable” or “cannot answer” are classified as abstentions.

3.3 RAG Model

To evaluate the efficacy of non-parametric knowledge integration, we implemented a Retrieval-Augmented Generation (RAG) system. Unlike fine-tuning, which encodes knowledge into model weights, RAG grounds generation in explicitly retrieved context. Our architecture consists of two primary components: a dense retriever and a generative reader.

Retrieval Component. We constructed a dense vector index over SQuAD 2.0 context passages to enable semantic search. We utilized multi-qa-MiniLM-L6-cos-v1 from the Sentence Transformers library (Reimers and Gurevych, 2019), which maps text into a 384-dimensional dense vector space optimized for semantic search and question-answering tasks. We implemented the index using FAISS (Facebook AI Similarity Search) (Johnson et al., 2017) with an IndexFlatL2 structure, performing exhaustive Euclidean distance searches to guarantee exact nearest-neighbor retrieval without approximation errors. The SQuAD 2.0 dataset was preprocessed to extract unique context paragraphs, yielding approximately 150,000 entries with no duplicates.

Generation Pipeline. The generation process follows a retrieve-then-generate workflow. For each query, the system retrieves the top $k = 5$ most relevant context paragraphs. We construct a prompt concatenating these retrieved passages with the same instruction used in the baseline, explicitly stating: “If the answer is not present in the provided context, you must abstain.” We use the same Llama-3.2-3B-Instruct base model as the baseline and fine-tuned experiments to ensure fair comparison.

Abstention Logic. To detect unanswerable questions, we implemented a two-stage filter. First, the model is instructed to output a specific abstention token. Second, a fallback mechanism checks if the answer generation fails to follow the extractive format, classifying such failures as “unanswerable”.

3.4 Fine-Tuned Model

To enable direct comparison with the baseline, we fine-tune the same Llama-3.2-3B-Instruct model using QLoRA (Dettmers et al., 2023), a parameter-efficient method that enables fine-tuning of quan-

tized models while preserving performance. We load the base model in 4-bit precision using NormalFloat4 (NF4) quantization with double quantization enabled, reducing memory footprint while maintaining model quality. Computation is performed in bfloat16 precision.

We apply low-rank adapters to both attention layers (q_proj, k_proj, v_proj, o_proj) and MLP layers (gate_proj, up_proj, down_proj), as the QLoRA paper demonstrates that targeting all linear layers improves adaptation quality compared to attention-only configurations. We use rank $r = 16$ with scaling factor $\alpha = 32$ (following the common heuristic of $\alpha = 2r$), and dropout of 0.05. This configuration yields 24.3M trainable parameters (0.75% of the 3.2B total), enabling efficient fine-tuning on a single GPU.

We construct training examples from the SQuAD 2.0 training split by pairing each question-context input with a target: the extracted answer span for answerable questions, or the literal string “unanswerable” for unanswerable questions. This formulation trains the model to both extract answers and recognize knowledge boundaries within a single generation objective. Due to computational constraints, we randomly subsample 8,000 examples from the training split (seed=42) rather than using the full 130K examples. We reserve 5% of the original training data (6,516 examples) for validation during training.

We train for one epoch to match the baseline’s zero-shot setting in terms of minimal task exposure. Training uses a batch size of 1 with gradient accumulation over 8 steps (effective batch size of 8), learning rate of 2×10^{-4} with linear decay, and maximum sequence length of 384 tokens. Following the QLoRA paper’s recommendation, we mask prompt tokens during loss computation (setting labels to -100), training only on the target completion. This focuses learning on answer generation rather than prompt reconstruction.

3.5 Evaluation Metrics

We evaluate model performance using four metrics that capture different aspects of QA quality. All metrics are computed separately for answerable and unanswerable questions to assess the accuracy-abstention tradeoff that motivates our comparison.

Exact Match (EM) measures whether the predicted answer exactly matches the ground truth string after normalization (lowercasing, punctuation removal, article removal). This provides a

strict correctness metric. For unanswerable questions, EM captures whether the model correctly outputs an abstention token. When multiple reference answers exist, we take the maximum EM score across all references.

F1 Score computes token-level precision and recall between prediction and ground truth, offering partial credit for incomplete answers. This metric is more forgiving than EM for extractive QA where answer boundaries may vary slightly. As with EM, we take the maximum F1 score across all reference answers.

BERTScore F1 (Zhang et al., 2020) evaluates semantic similarity between generated and reference answers using contextual embeddings from RoBERTa-large with baseline rescaling. This metric assesses whether answers are semantically equivalent even when not lexically identical, helping identify paraphrasing versus hallucination. For examples with multiple references, we select the reference that maximizes BERTScore F1 and report the corresponding precision, recall, and F1 values.

Semantic Textual Similarity (STS) measures cosine similarity between sentence embeddings of the generated answer and reference answer using the all-mpnet-base-v2 model (Reimers and Gurevych, 2019). We normalize similarity scores to the $[0,1]$ range using $(s+1)/2$. High STS scores indicate answers are semantically aligned with reference answers. For multiple references, we take the maximum similarity score. This metric helps quantify whether answers derive from the given context rather than parametric memory, directly addressing our research question about external versus internal knowledge integration.

For abstained predictions (unanswerable outputs or empty generations), we use the placeholder text "[NO ANSWER]" for BERTScore and STS computations to ensure consistent comparison with unanswerable ground truth labels.

4 Results and Discussion

Table 3 presents the performance comparison between the zero-shot baseline Llama-3.2-3B-Instruct model, the RAG-based system, and the QLoRA fine-tuned variant across all evaluation metrics.¹

¹Baseline and RAG metrics are computed on the full SQuAD 2.0 validation set (11,873 examples), while QLoRA metrics are from a 2,000-example subset. Future work will evaluate all models on identical subsets for stricter comparison.

4.1 Comparing Baseline with RAG

We evaluated both the Zero-Shot Baseline (Llama 3B) and the RAG System on the SQuAD 2.0 validation set. The analysis focuses on three key areas: overall performance, the ability to answer correctly when possible, and the ability to abstain when necessary.

The most significant finding is that the RAG system is significantly superior at handling unanswerable questions (+7.60% improvement in EM). The RAG system acts "safely": when it cannot find a semantic match in the retrieved documents, it correctly abstains nearly 60% of the time.

However, the RAG system suffered a significant performance drop on Answerable questions (22.22% EM vs. Baseline's 38.36%). This highlights a critical "retrieval bottleneck." Even when the context window was expanded (from $k=3$ to $k=5$), the system frequently failed to generate the correct answer. This suggests that when the retriever fails to surface the exact evidence required, the LLM is forced to abstain due to strict prompt instructions, whereas the Baseline can leverage its internal training data to answer correctly.

Additionally, the Baseline achieved a higher BERTScore F1 (45.35%) compared to RAG (38.57%). This indicates that when the Baseline generates an answer, it is semantically closer to the ground truth. The lower score for RAG implies that when it does attempt to answer (rather than abstaining), the retrieved context might occasionally mislead the generation.

In summary, the RAG architecture successfully met the project's primary objective of improving unanswerable question detection. By explicitly grounding the generation in retrieved context, we reduced the rate of hallucination on unanswerable queries compared to the baseline. However, this came at the cost of sensitivity; the system became over-cautious, frequently abstaining on questions it could have answered, simply because the retrieval step was imperfect.

4.2 Comparing Baseline with Q-Lora

The QLoRA fine-tuned model reveals a critical failure mode in teaching LLMs to recognize knowledge boundaries through low-resource fine-tuning. While fine-tuning achieved the highest unanswerable detection score—Exact Match rose from 52.01% to 74.53%, with BERTScore F1 increasing from 44.26% to 67.50%—this improve-

Table 3: Baseline vs. RAG vs. QLoRA fine-tuned model performance on SQuAD 2.0.

| Metric | Baseline 3B | | RAG 3B | | QLoRA 3B | |
|--------------|-------------|--------|--------|--------|----------|--------|
| | Ans. | Unans. | Ans. | Unans. | Ans. | Unans. |
| Exact Match | 38.36 | 52.01 | 22.22 | 59.61 | 2.99 | 74.53 |
| F1 Score | 59.58 | 52.01 | 39.41 | 59.61 | 18.93 | 74.53 |
| BERTScore F1 | 46.45 | 44.26 | 24.23 | 52.60 | 10.08 | 67.50 |
| STS Score | 85.83 | 78.41 | 76.50 | 81.87 | 71.14 | 88.38 |

ment was not due to learned discrimination but rather because the model learned abstention as a dominant strategy. This behavior is evidenced by the catastrophic collapse in answerable question performance: Answerable Exact Match fell from 38.36% (Baseline) to just 2.99% (QLoRA). Similarly, the F1 score for answerable questions dropped to 18.93%. The Semantic Textual Similarity (STS) scores reinforce this pattern: while the model achieved high similarity on unanswerable questions (88.38%), this reflects a tendency to output “unanswerable” regardless of the input. This asymmetry suggests that within our constrained training regime (8,000 randomly sampled examples, single epoch), the model resorted to the shortcut of always abstaining, rendering it functionally useless for actual question answering. The finding highlights a key challenge for fine-tuning approaches: optimizing for unanswerable question detection may inadvertently suppress the extractive capabilities that make QA systems useful, a trade-off that retrieval-augmented approaches may handle differently by externalizing knowledge boundaries.

5 Conclusion

We presented a systematic comparison of RAG and fine-tuning for unanswerable question detection, evaluating how different knowledge integration paradigms affect abstention behavior in generative LLMs. Using SQuAD 2.0 and a controlled experimental setup with Llama-3.2-3B-Instruct, we isolated the effect of the knowledge integration approach while holding model capacity constant.

Key findings. Both RAG and QLoRA fine-tuning improve unanswerable question detection over the zero-shot baseline, but exhibit distinct accuracy-abstention tradeoffs. QLoRA achieves the highest abstention performance (74.53% EM on unanswerable questions) but suffers catastrophic degradation on answerable questions (2.99% EM). RAG offers more balanced performance, improving unanswerable detection to 59.61% EM while re-

taining moderate answerable performance (22.22% EM). These results suggest that externalizing knowledge boundaries through retrieval yields more graceful degradation than encoding them parametrically, where the model learns abstention as a dominant strategy.

Limitations. Our findings are subject to several constraints. The QLoRA model was trained on only 6% of available training data for a single epoch due to computational limitations, which likely contributed to the over-abstention behavior. The QLoRA evaluation used a smaller validation subset than the baseline and RAG experiments, limiting direct comparability. Additionally, our study uses a single model architecture (Llama 3B) and dataset (SQuAD 2.0); generalization to other models, scales, and domains remains to be validated.

Future work. Several directions emerge from our findings. First, training QLoRA on the full dataset with stratified sampling or weighted loss functions may yield more balanced abstention behavior. Second, hybrid approaches that combine retrieval’s external grounding with fine-tuning’s learned judgment warrant exploration. Third, multi-task objectives that explicitly balance answer extraction and abstention detection could mitigate the tradeoff we observed. Finally, extending this comparison to domain-specific settings (e.g., medical or legal QA) would assess whether our findings hold in the high-stakes applications that motivated this work.

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 Abstract, Introduction, and Background sections. Victoria was also the primary author for the Data Description, Baseline Model, and Evaluation Metrics subsections of the Methods section.

Alejandro Franza contributed the end-to-end implementation of the Retrieval-Augmented Gen-

eration (RAG) system. Alejandro was the primary author for the RAG Model Design and Implementation subsection of the Methods section and the Comparing Baseline with RAG subsection within Results and Discussion.

Eric Tsang contributed the QLoRA fine-tuning implementation and the comparison analysis between the baseline and QLoRA performance. Eric was the primary author of the Fine-Tuned Model, Comparing Baseline with Qlora, and the Conclusion section.

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.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#).
- 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).
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. [Billion-scale similarity search with gpus](#). *IEEE Transactions on Big Data*, 7:535–547.
- 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.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 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. [DelusionQA: 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 Hasiibi. 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 Baseline Prompt

A.1 Test Data Construction

We sampled 2,000 examples from the SQuAD 2.0 training set, stratified by:

Answerability: Answerable vs. unanswerable questions (equal representation).

Complexity: Defined as the average of standardized token counts for context, question, and answer. We partition into three levels (low, medium, high) using quantile binning, resulting in six strata. Proportional sampling ensures balanced representation across all combinations.

A.2 Instruction Variants

V0 (Simple):

Answer the question based solely on the given context. If the answer is not in the context, respond with 'unanswerable'.

V1 (Explicit):

You are an extractive question answering system. Extract the shortest possible answer span from the context that fully answers the question. Extract the answer word-for-word from the context. Do not paraphrase or generate new text. Do not use outside knowledge. Do not provide an explanation of your response. Just the answer span. If the question cannot be answered from the context, output exactly: 'unanswerable'.

V2 (Chain-of-Thought):

You are an extractive question answering system. Answer the question based on the context given. Think step by step: 1. First, determine if the context contains information to answer the question. 2. If yes, extract the exact text span that answers it. 3. If no, respond with just 'unanswerable'.

Table 4: Instruction comparison on test set.

| Instruction | EM | F1 | BERTScore | STS |
|---------------|-------|-------|-----------|-------|
| V0 (Simple) | 45.25 | 56.28 | 43.22 | 80.70 |
| V1 (Explicit) | 43.55 | 54.94 | 43.23 | 81.47 |
| V2 (CoT) | 34.05 | 46.62 | 33.16 | 77.51 |

A.3 Results

Table 4 shows performance across variants using temperature=0.1, top_p=0.9, max_new_tokens=35.

V1 (Explicit) achieves the highest BERTScore F1 (43.23%), demonstrating superior semantic alignment. We select this variant for its emphasis on extractive behavior and clear abstention guidelines.