Fine-Tuning vs. RAG for Multi-Hop Question Answering with Novel Knowledge

Zhuoyi Yang[1] **Yurun Song**[1] **Kyler Harris**[2] **Iftekhar Ahmed**[1] **Ian G. Harris**[1]
[1]University of California, Irvine, Irvine, California, USA
[2]University of California, Santa Cruz, Santa Cruz, California, USA
{zhuoyy1,yuruns,iftekha,iharris}@uci.edu
kgharris@ucsc.edu

Abstract

Multi-hop question answering is widely used to evaluate the reasoning capabilities of large language models (LLMs), as it requires integrating multiple pieces of supporting knowledge to arrive at a correct answer. While prior work has compared fine-tuning and retrieval-augmented generation (RAG) for factual recall and single-hop question answering, it remains unclear how these approaches perform in multi-hop settings that require compositional reasoning over temporally novel knowledge. In particular, prior comparisons often do not control for model scale, evaluation format, or knowledge freshness, making it difficult to isolate the effect of knowledge injection mechanisms.

In this paper, we systematically compare parametric and non-parametric knowledge injection methods for open-domain multi-hop question answering. We evaluate unsupervised fine-tuning (continual pretraining), supervised fine-tuning, and retrieval-augmented generation across three 7B-parameter open-source LLMs. Experiments are conducted on two benchmarks: Question Answering Science Challenge (QASC), a standard multi-hop science question answering dataset, and a newly constructed dataset of over 10,000 multi-hop questions derived from Wikipedia events in 2024, which is designed to test knowledge beyond the models' pretraining cutoff.

Our results show that unsupervised fine-tuning provides only limited gains over base models, suggesting that continual pretraining alone is insufficient for improving multi-hop reasoning accuracy. In contrast, RAG yields substantial and consistent improvements, particularly when answering questions that rely on temporally novel information. Supervised fine-tuning achieves the highest overall accuracy across models and datasets. These findings highlight fundamental differences in how knowledge injection mechanisms support multi-hop question answering and underscore the importance of retrieval-based methods when external or compositional knowledge is required.

1 Introduction

Large language models (LLMs) have demonstrated strong performance on a wide range of question answering tasks, including those that require combining information from multiple sources (Yang et al., 2018). Among these tasks, multi-hop question answering has emerged as a key benchmark for evaluating a model's ability to integrate multiple pieces of evidence and perform compositional reasoning (Ho et al., 2020). Rather than relying on a single fact, multi-hop questions require models to connect several supporting facts—often drawn from different documents or distant parts of a text—to arrive at the correct answer (Trivedi et al., 2022).

One key factor that influences multi-hop question answering performance is the mechanism used to inject knowledge into the model. Broadly, existing approaches fall into two categories. Parametric methods incorporate knowledge directly into model parameters through weight updates, either using unlabeled text (unsupervised fine-tuning or continual pretraining) (Gururangan et al., 2020) or labeled question–answer pairs (supervised fine-tuning) (Khashabi et al., 2020). In contrast, non-parametric methods such as retrieval-augmented generation (RAG) provide external evidence at inference time by retrieving relevant documents from a knowledge corpus and conditioning the model's predictions on the retrieved text (Lewis et al., 2020).

Prior work has compared fine-tuning and RAG primarily in the context of factual recall and single-hop question answering, often finding retrieval-based methods to be competitive or superior when external knowledge is required (Ovadia et al., 2024). However, most prior comparisons focus on single-hop question answering or factual recall, where answers can often be obtained from a sin-

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 384–392

gle piece of evidence. It remains unclear whether these findings extend to multi-hop question answering, which requires identifying multiple relevant facts and composing them coherently (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022). Moreover, existing studies do not explicitly examine settings involving temporally novel knowledge beyond the model's pretraining cutoff. They also do not control for confounding factors such as model scale and evaluation format. As a result, the relative effectiveness of parametric and non-parametric knowledge injection mechanisms for multi-hop reasoning remains insufficiently understood. In such settings, retrievers may fail to retrieve all necessary evidence, while frozen LLMs may struggle to effectively integrate retrieved information through in-context learning alone (Ho et al., 2020; Yang et al., 2018; Trivedi et al., 2022).

In this paper, we aim to answer the following research question (RQ): Does the method of knowledge injection impact the effectiveness of open-domain multi-hop question answering? To address this question, we conducted a systematic comparison of fine-tuning and RAG under controlled experimental settings. We consider unsupervised fine-tuning, supervised fine-tuning, and RAG as three distinct knowledge injection mechanisms that differ in when and how supporting knowledge is made available to the model.

We evaluate these methods on two multi-hop benchmarks that represent different knowledge conditions. The first benchmark is QASC dataset (Khot et al., 2020), a widely used multi-hop science question answering dataset. The second benchmark is 2024 Events dataset, a newly constructed dataset of over 10,000 multi-hop questions derived from Wikipedia events in 2024,[1] designed to assess performance under temporally novel knowledge that is not memorized during pretraining. Using three open-source 7B-parameter LLMs, we compare answer accuracy across all settings using a unified multiple-choice evaluation framework.

Our experimental results show consistent trends across both datasets. Unsupervised fine-tuning yields only marginal improvements with the base models, suggesting that continual pretraining alone is insufficient for improving multi-hop reasoning accuracy. RAG substantially improves performance, more than doubling accuracy on the 2024 Events dataset. Supervised fine-tuning achieves the highest overall accuracy, highlighting the strong effect of task-specific supervision. **We make the following contributions:**

- We provide the first controlled comparison of parametric (unsupervised and supervised fine-tuning) and non-parametric (RAG) knowledge injection methods specifically for multi-hop question answering, isolating the effect of knowledge access while holding model scale and evaluation format constant.
- We introduce a new multi-hop question answering benchmark based on 2024 Wikipedia events, enabling evaluation under temporally novel knowledge conditions beyond the models' pretraining cutoff.
- We show that different knowledge injection mechanisms offer different strengths: retrieval-augmented generation is particularly effective for questions requiring access to novel external knowledge, while supervised fine-tuning yields the highest overall accuracy when task-specific labeled data is available.

We will release the source code upon acceptance of this paper.

2 Related Work

2.1 Multi-hop Question Answering as a Reasoning Benchmark

Multi-hop question answering has been widely adopted as a benchmark for evaluating a model's ability to process information from multiple sources. Datasets such as QASC, HotpotQA, 2WikiMultiHopQA, and MuSiQue are specifically designed to require the aggregation of evidence across multiple facts, sentences, or documents. Because these tasks cannot typically be solved using a single retrieved fact, performance on multi-hop benchmarks is often interpreted as a proxy for a model's reasoning capability (Yang et al., 2018; Trivedi et al., 2022; Khot et al., 2020; Ho et al., 2020).

Most prior work evaluates multi-hop question answering systems using answer accuracy as the primary metric, without explicitly supervising or inspecting intermediate reasoning steps (Yang et al., 2018). As a result, improvements in accuracy are often attributed to better reasoning, although they

[1] https://en.wikipedia.org/wiki/Category:2024_in_the_United_States_by_month

may also reflect gains in retrieval quality, memorization, or task-specific heuristics (Bender et al., 2021). This limitation has motivated recent studies to examine how architectural choices, prompting strategies, and training objectives influence multi-hop performance, even when explicit reasoning traces are not available (Wei et al., 2022).

Our work follows this evaluation paradigm by using accuracy as the primary metric (Ovadia et al., 2024), but focuses specifically on how different knowledge injection mechanisms affect multi-hop question answering outcomes. Rather than proposing a new reasoning architecture or dataset, we study how models behave under different ways of accessing supporting knowledge, holding model scale and evaluation format constant.

2.2 Knowledge Injection in Large Language Models

Large language models acquire substantial factual knowledge during pretraining (Petroni et al., 2019), but this knowledge is inherently static and bounded by the training corpus (Brown et al., 2020). To address these limitations, prior work has explored various mechanisms for injecting additional knowledge into LLMs, including continual pretraining (Gururangan et al., 2020), supervised fine-tuning (Khashabi et al., 2020), and retrieval-based methods (Jiang et al., 2023b; Asai et al., 2023; Wang et al., 2025). These approaches differ fundamentally in whether knowledge is encoded parametrically in model weights or provided dynamically at inference time.

Unsupervised fine-tuning, also referred to as continual pretraining, incorporates new information by further training a model on unlabeled text using a language modeling objective. This approach is appealing due to its scalability and lack of annotation requirements. However, prior studies have shown that continual pretraining struggles to internalize sparse or temporally novel facts, and its benefits for downstream reasoning tasks are often limited (Lewis et al., 2020; Gururangan et al., 2020). In the context of multi-hop question answering, this method requires the model to implicitly encode all relevant supporting knowledge within its parameters prior to inference.

Supervised fine-tuning injects knowledge through labeled examples, typically in the form of question–answer pairs or instruction-following data. This approach has been shown to substantially improve downstream task performance and alignment with evaluation formats (Brown et al., 2020; Ouyang et al., 2022). However, supervised fine-tuning may conflate knowledge acquisition with task-specific pattern learning, making it difficult to disentangle improvements due to reasoning ability from those due to answer format adaptation or shortcut learning (Geirhos et al., 2020). While widely used, supervised fine-tuning does not directly address how models access or combine external knowledge at inference time.

Retrieval-augmented generation (RAG) represents a non-parametric alternative, in which relevant documents are retrieved from an external corpus and provided to the model as additional context during inference (Lewis et al., 2020). By decoupling knowledge storage from model parameters, RAG enables access to large and dynamically updated knowledge sources and has been shown to be effective for factual recall and knowledge-intensive tasks. Several studies have demonstrated strong RAG performance on single-hop or weakly compositional question answering tasks. However, less is known about how retrieval quality and context integration affect performance on multi-hop questions that require composing multiple pieces of evidence (Ovadia et al., 2024; Jiang et al., 2023b; Asai et al., 2023; Wang et al., 2025).

Our work builds on this literature by directly comparing unsupervised fine-tuning, supervised fine-tuning, and retrieval-augmented generation under a unified evaluation framework. Unlike prior studies that focus primarily on single-hop factual recall or open-ended generation, we examine how these knowledge injection mechanisms influence multiple-choice multi-hop question answering, including settings involving temporally novel knowledge.

3 Overview

Figure 1 provides an overview of our experimental framework for studying multi-hop question answering under different knowledge injection mechanisms. This figure highlights the key difference between parametric and non-parametric knowledge injection at training vs. inference time. To enable a controlled comparison, we construct both the **knowledge base** and **benchmark datasets** consisting of multi-hop questions with corresponding answers. We consider three representative knowledge injection strategies: **unsupervised fine-tuning, supervised fine-tuning, and retrieval-augmented**

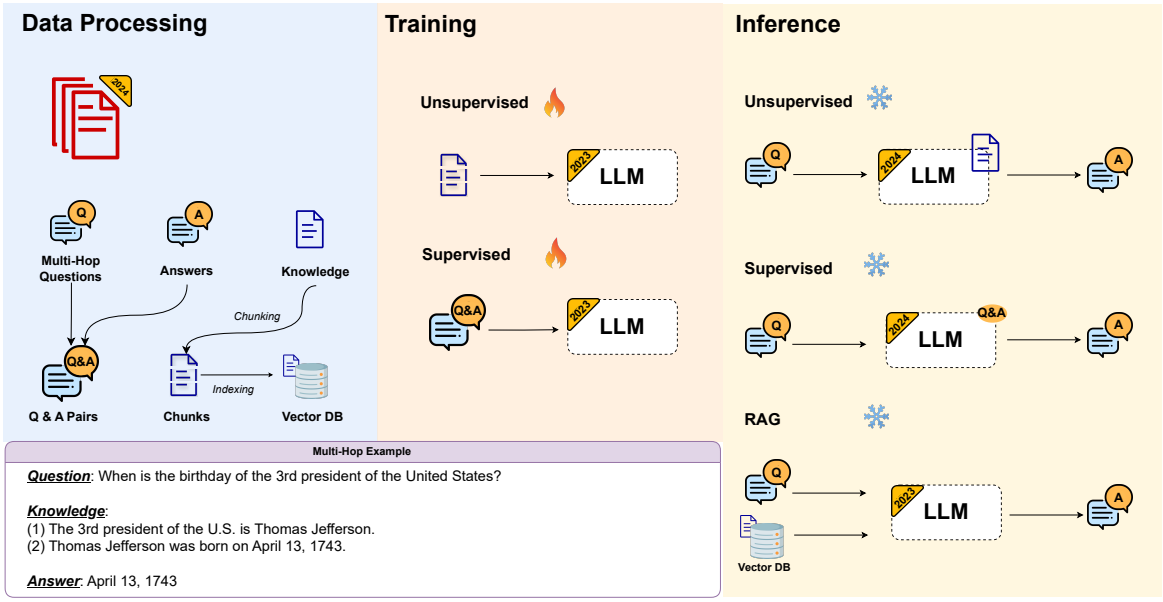

Figure 1: Comparative Framework for Multi-Hop Question Answering across Knowledge Injection Methods

generation (RAG).

For unsupervised fine-tuning, knowledge is incorporated into the model via unlabeled text. For supervised fine-tuning, knowledge is embedded into the model through question-answer pairs. In contrast, RAG injects external knowledge at inference time by retrieving relevant evidence from an external corpus. Using a unified multiple-choice evaluation setup, we assess the performance of all three mechanisms in terms of answer accuracy on the same benchmarks.

We explicitly control whether models are exposed to temporally novel information during training and inference. We construct a 2024-specific Wikipedia knowledge corpus and use it for unsupervised fine-tuning, supervised fine-tuning, and as the external corpus for RAG. This design allows us to compare parametric and non-parametric knowledge injection under post–pretraining-cutoff knowledge and to evaluate the effectiveness of three knowledge injection mechanisms on novel knowledge.

A key design goal of our framework is to isolate the effect of when and how knowledge is injected into the model, while controlling for model architecture, scale, and evaluation format. By evaluating all methods using a unified multiple-choice scoring procedure, we minimize confounding effects arising from generation style, decoding strategies, or output formatting. This allows differences in accuracy to be more directly attributed to the underlying knowledge access mechanism rather than surface-level task alignment.

4 Knowledge Base Construction

4.1 Benchmark Selection and Rationale

We selected QASC Dataset and created our own 2024 Events Dataset from Wikipedia content.

QASC Dataset (Khot et al., 2020) QASC (Question Answering via Sentence Composition) is a widely-used multi-hop reasoning dataset. Each question has eight answer options, among which only one is correct. We chose this dataset because it includes questions from different fields of science, which enables a more thorough evaluation of the model's reasoning capability. However, we acknowledge that answers to questions in this dataset may have been partially observed during pre-training.

2024 Events Dataset We constructed a dataset of over 10,000 multi-hop multiple-choice questions from events that happened in 2024 based on Wikipedia content[1]. Each question has four answer options with only one being correct. We describe how this dataset is constructed in details in Section 4.2. This design enables the evaluation of multi-hop question answering under temporally novel knowledge conditions. Since these articles postdate the training cutoff of our evaluated models, they are not memorized by the models during pre-training.

4.2 Data Collection and Knowledge Corpora

QASC Dataset Following prior work (Khashabi et al., 2018), we categorized questions in this dataset into seven scientific domains: Physics, Chemistry, Biology, Earth Science, Astronomy, Environmental Science, and General Science. We use GPT-4 (Achiam et al., 2024) to label each question accordingly.

We used the external knowledge corpus provided with the QASC dataset (Khot et al., 2020), which consists of approximately 17 million science-related sentences across multiple domains. This fixed corpus was built prior to evaluation and serves as both unsupervised fine-tuning training corpus and the retrieval source.

2024 Events Dataset For the 2024 events dataset, we constructed the knowledge corpus from scratch using Wikipedia. We used the Wikipedia category page "2024 in the United States by month[1]" as a seed and follow links to individual monthly pages. From each month, we extracted descriptions of real-world events that occurred during that period. All the events together formed our knowledge base.

To create multi-hop questions, we segmented each event description into chunks of 200 tokens to ensure manageable context lengths. For each chunk, we then prompted GPT-4 and DeepSeek-R1(DeepSeek-AI et al., 2025) to independently generate two multi-hop multiple-choice questions, resulting in a diverse set of questions grounded in temporally novel information.

5 Experimental Setup

5.1 Model Selection

We evaluate three open-source large language models with approximately 7 billion parameters: Mistral-7B, LLaMA-7B, and LLaMA-7B-Instruct (Jiang et al., 2023a; Touvron et al., 2023a,b). Mistral-7B and LLaMA-7B are base pretrained models that have not undergone instruction tuning, and therefore provide a view of multi-hop question answering performance without explicit reasoning- or instruction-oriented alignment. In contrast, LLaMA-7B-Instruct has been instruction-tuned to better follow natural language prompts and perform reasoning-style tasks, and serves as a stronger upper bound on achievable performance. Using models with comparable parameter counts allows us to control for model scale while examining the impact of different knowledge access mechanisms on multi-hop question answering accuracy. All three models have pre-training data cutoff point before 2024.

5.2 Implementation Details

5.2.1 Parametric Injection

Common Setup All experiments were run on two NVIDIA A6000 GPUs. Unless otherwise stated, we initialized from the base model and apply LoRA to the query and value projections of each self-attention layer ($r = 16, \alpha = 32, dropout = 0.1$), freezing all base parameters. We optimized with AdamW ($lr = 1e-5$) using a linear decay schedule with 10% warmup, trained for 20 epochs with bfloat16 mixed precision, batch size 16.

Unsupervised Fine-tuning We performed causal language modeling on raw text. For QASC dataset, we used the released external knowledge corpus (17M science sentences) and trained the model at the sentence level. For the 2024 Events dataset, we scraped Wikipedia event articles as described in Section 4.2 and segmented them into 1,000-token chunks with 200-token overlap. The model was trained using the standard autoregressive language modeling objective. Given a token sequence $(x_1, \ldots, x_T)$, we minimize:

$$\mathcal{L}_{\text{CLM}} = -\sum_{t=1}^{T} \log p(x_t \mid x_{<t}). \tag{1}$$

Supervised Fine-tuning We finetuned the model on multiple-choice question answering by concatenating the question with each candidate answer option and predicting the correct option using a classification head. We considered $C \in \{4, 8\}$ answer options (QASC: $C = 8$; 2024 Events: $C = 4$). For the 2024 Events dataset, we used DeepSeek-generated questions for training and GPT-generated questions for evaluation. The training objective minimizes the cross-entropy loss over the C answer classes:

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp(z_y)}{\sum_{c=1}^{C} \exp(z_c)}, \tag{2}$$

where $z \in \mathbb{R}^C$ denotes the output logits and $y \in \{1, \ldots, C\}$ is the index of the correct answer option.

5.2.2 RAG Pipeline

Knowledge Corpus and Indexing We constructed a knowledge corpus from a 2024 Wikipedia text

Task	Model	Base model	Base model + RAG	Unsupervised Fine-tuned	Supervised Finetuned
Physics	Mistral-7B	0.354	0.634	0.362	0.806
	Llama2-7B	0.383	0.659	0.372	0.812
	Llama2-7B-Instruct	0.457	0.757	0.449	0.864
Chemistry	Mistral-7B	0.345	0.623	0.356	0.812
	Llama2-7B	0.351	0.642	0.343	0.823
	Llama2-7B-Instruct	0.392	0.735	0.413	0.874
Biology	Mistral-7B	0.353	0.602	0.367	0.801
	Llama2-7B	0.359	0.614	0.362	0.818
	Llama2-7B-Instruct	0.401	0.717	0.409	0.862
Earth Science	Mistral-7B	0.353	0.604	0.359	0.812
	Llama2-7B	0.367	0.615	0.379	0.826
	Llama2-7B-Instruct	0.405	0.748	0.413	0.861
Astronomy	Mistral-7B	0.362	0.598	0.375	0.793
	Llama2-7B	0.381	0.634	0.394	0.791
	Llama2-7B-Instruct	0.410	0.739	0.427	0.854
Environmental Science	Mistral-7B	0.341	0.595	0.356	0.803
	Llama2-7B	0.351	0.613	0.369	0.813
	Llama2-7B-Instruct	0.403	0.735	0.412	0.868
General Science	Mistral-7B	0.352	0.641	0.365	0.783
	Llama2-7B	0.351	0.654	0.372	0.815
	Llama2-7B-Instruct	0.400	0.759	0.414	0.864

Table 1: Results for the QASC dataset in terms of accuracy

	Base Model	Base Model + RAG	Unsupervised Fine-tuning	Supervised Fine-tuning
Mistral-7B	0.276	0.654	0.320	0.813
Llama2-7B	0.278	0.672	0.329	0.832
Llama-7B-Instruct	0.326	0.753	0.392	0.884

Table 2: Results for the 2024 Events dataset in terms of accuracy

dump as described in Section 4.2. Articles are segmented into overlapping chunks using a recursive character-based splitter (Pandya and Holia, 2023) with a chunk size of 1000 tokens and an overlap of 300 tokens. Each chunk is embedded using the BGE-large-en dense embedding model(Xiao et al., 2024) and indexed with FAISS using inner-product similarity over L2-normalized vectors (Johnson et al., 2019).

Retrieval and Reranking Given a question, we retrieve the top 20 candidate chunks using dense retriever BGE. These candidates are then reranked using a cross-encoder reranker (ms-marco-MiniLM-L6-v2(Pande et al., 2025)), and the top 4 chunks are selected as contextual evidence.

Prompt Construction The retrieved chunks are concatenated and prepended to the question and its answer options to form a single prompt. The model is instructed to answer using only the provided context and to output a single answer letter. To improve alignment with the desired behavior, we include three in-context examples in the prompt.

Answer Scoring and Selection Rather than generating free-form text, we adopt an MMLU-style scoring strategy (Hendrycks et al., 2021): Given the prompt, we compute the log-probability of generating each answer option token at the next position. The option with the highest log-probability is selected as the final prediction. This approach ensures direct comparability with classification-based evaluation while avoiding ambiguity from open-ended generation.

5.3 Evaluation Method

All models were evaluated under a unified multiple-choice classification framework, where the goal was to select the only correct answer from option candidates. We used accuracy as the primary evaluation metric. Accuracy is defined as the fraction of questions for which the predicted answer label matches the ground-truth label:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\hat{y}_i = y_i], \qquad (3)$$

where N denotes the number of evaluation examples. The primary difference across settings lies in how the score for each option is computed.

5.3.1 Supervised finetuned models

For supervised fine-tuning experiments, models were implemented using a sequence classification head with $|\mathcal{C}|$ output logits, where $\mathcal{C}$ denotes the set of answer labels (e.g., $\{A, B, C, D\}$).

Given an input example, the model produces logits $z \in \mathbb{R}^{|\mathcal{C}|}$. The predicted answer is selected as:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} z_c. \tag{4}$$

Evaluation accuracy is computed by comparing the predicted label $\hat{y}$ with the gold label y over the held-out test set.

5.3.2 Unsupervised Finetuned models and RAG pipeline

Rather than generating free-form text, we adopted an MMLU-style label scoring strategy. Given the prompt, we computed the conditional log-probability of each answer option label. At the next-token position, the predicted answer corresponds to the option with the highest log-probability:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \log p(c \mid \text{prompt}), \tag{5}$$

where $\mathcal{C}$ denotes the set of answer labels. $\mathcal{C} = \{A, B, C, D\}$ for 2024 Wikipedia Events dataset and $\mathcal{C} = \{A, B, C, D, E, F, G, H\}$ for QASC dataset.

6 Results

For each task and model, we compare four approaches: the base model, RAG using the same base model as the generator, unsupervised fine-tuning, and supervised fine-tuning.

6.1 QASC Dataset

The results on the QASC dataset are reported in Table 1. Across all tasks, the base models achieve substantially lower accuracy than both RAG and supervised fine-tuning. Nevertheless, the base models consistently outperform random guessing, which yields an expected accuracy of $\frac{1}{C}$ (i.e., $\frac{1}{8}$ for QASC). This indicates that the models retain some relevant knowledge from pretraining.

Unsupervised fine-tuning yields only marginal improvements across all three models, suggesting that continual pretraining alone provides limited benefits for multi-hop reasoning. In contrast, RAG improves accuracy by approximately 30 percentage points across all models. Supervised fine-tuning achieves the highest overall accuracy on this dataset.

6.2 2024 Events Dataset

The evaluation results on the 2024 Events dataset are summarized in Table 2. Similar trends are observed across all models. Base models perform poorly on this dataset, reflecting the difficulty of answering questions that rely on temporally novel knowledge. The instruction-tuned model achieves accuracy notably above random guessing, which corresponds to $\frac{1}{C}$ (i.e., $\frac{1}{4}$), despite lacking explicit access to the underlying event knowledge.

Unsupervised fine-tuning again yields only marginal improvements. In contrast, RAG more than doubles accuracy across all models, highlighting the effectiveness of retrieval-based methods when external, up-to-date knowledge is required. Supervised fine-tuning achieves the largest performance gains overall.

7 Conclusion

In this paper, we presented a systematic comparison of parametric and non-parametric knowledge injection mechanisms for open-domain multi-hop question answering. Through controlled experiments on both the QASC benchmark and a newly constructed 2024 Events dataset designed to test temporally novel knowledge, we showed that unsupervised fine-tuning via continual pretraining yields only marginal gains on answer accuracy, suggesting limited effectiveness for improving multi-hop reasoning. In contrast, RAG substantially improves performance, particularly in settings where required knowledge lies beyond the model's pre-training cutoff, while supervised fine-tuning achieves the highest overall accuracy when task-specific labeled data is available. These findings highlight fundamental differences in how models access and utilize knowledge under different injection strategies and underscore the importance of retrieval-based methods for reasoning over novel or compositional information. Together, our results provide practical guidance for selecting knowledge injection approaches when deploying large language models for multi-hop question answering tasks.

8 Ethical Statement

The knowledge corpus is derived from Wikipedia, which may reflect geographic and cultural biases present in its content. The use of LLMs for dataset construction introduces the risk of hallucinated or incorrect information, which may affect evaluation reliability. Training and evaluating large language models incur computational and environmental costs. Future work should consider more efficient methods.

9 Limitations

While our study provides a controlled comparison of different knowledge injection mechanisms, several limitations remain. First, our evaluation focuses on answer accuracy, which may not fully capture the complexity of multi-hop reasoning processes. Second, the effectiveness of retrieval-augmented generation may vary depending on the quality of the retrieval component and the underlying knowledge corpus. Third, although we construct a dataset to evaluate temporally novel knowledge, it may not fully represent the diversity of real-world information-seeking scenarios. Finally, our experiments are conducted on a specific set of model scales and evaluation settings, and further investigation is needed to assess the generality of these findings across broader configurations.

References

John Achiam and 1 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

DeepSeek-AI and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in large language models. *arXiv preprint*, arXiv:2501.12948.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. In *Nature Machine Intelligence*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online).

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, and 1 others. 2023a. Mistral 7b. *arXiv preprint*, arXiv:2310.06825.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. *Preprint*, arXiv:2305.06983.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Question answering as global reasoning over semantic abstractions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *Preprint*, arXiv:1910.11473.

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 *Proceedings of the Advances in Neural Information Processing Systems*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in llms. *Preprint*, arXiv:2312.05934.

Manu Pande, Shahil Kumar, and Anay Yatin Damle. 2025. When fine-tuning fails: Lessons from ms marco passage ranking. *Preprint*, arXiv:2506.18535.

Keivalya Pandya and Mehfuza Holia. 2023. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations. *Preprint*, arXiv:2310.05421.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, arXiv:2307.09288.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Liang Wang, Haonan Chen, Nan Yang, Xiaolong Huang, Zhicheng Dou, and Furu Wei. 2025. Chain-of-retrieval augmented generation. *Preprint*, arXiv:2501.14342.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. *Preprint*, arXiv:2309.07597.

Zhiyuan Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.