The Small-Model Graph-RAG Gap Is a Faithfulness Gap

Jason Thomo
University of Victoria
Victoria, Canada
jthomo@uvic.ca

Venkatesh Srinivasan
Santa Clara University
Santa Clara, United States
vsrinivasan4@scu.edu

Alex Thomo
University of Victoria
Victoria, Canada
thomo@uvic.ca

Abstract

On multi-hop benchmarks, modern Graph-RAG retrieves the gold answer for 77–90% of questions, yet a budget open-weight 8B model answers only 31–77% correctly. We show that *faithfulness* (reading the retrieved evidence rather than guessing) and *calibration* (abstaining when it is insufficient) account for a large, recoverable part of this gap. A leak-controlled decomposition finds that on the hardest, least-contaminated benchmark (MuSiQue), 43% of questions have their supporting facts in the retrieved context yet are answered wrong, and on these the model gives a confident answer 82% of the time rather than abstaining. A grounding-and-calibration prompt makes the model answer from the retrieved context or abstain when it cannot, a more trustworthy operating point at no accuracy cost. A re-ask step then spends some of that calibration for coverage, turning abstentions into correct answers for a gain of +9 pp on KET-RAG and +13 pp on LightRAG (MuSiQue, answer-present slice) that closes about half the gap to a 70B model on KET-RAG and most of it on LightRAG. Few-shot example *source* moves the model along the same faithfulness-accuracy frontier, better for calibration than accuracy. Unlike structure-exploiting augmentations that degrade on uniform-relevance retrieval, grounding helps *most* there. For small-model Graph-RAG the target is faithful, calibrated use of retrieved evidence. Code is available at https://github.com/thomouvic/graphrag-faithfulness.

CCS Concepts

• **Computing methodologies** → **Natural language processing**;
• **Information systems** → **Question answering**.

Keywords

Graph-RAG, multi-hop question answering, faithfulness, calibration, selective prediction, few-shot prompting, knowledge graphs

ACM Reference Format:
Jason Thomo, Venkatesh Srinivasan, and Alex Thomo. 2026. The Small-Model Graph-RAG Gap Is a Faithfulness Gap. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3799682.3839910

1 Introduction

Multi-hop question answering requires combining several pieces of evidence scattered across a document collection. Graph-augmented retrieval-augmented generation (Graph-RAG) [2] addresses this by building a knowledge graph from the corpus and retrieving a connected, multi-hop context for a language model to read. On standard benchmarks (HotpotQA [24], MuSiQue [18]) state-of-the-art Graph-RAG, such as KET-RAG [8] and LightRAG [6], retrieves well. The gold answer is somewhere in the retrieved context for 77–90% of questions, yet a budget open-weight 8B model, the kind a cost-sensitive or private deployment would run, answers only 31–77% of them correctly [26]. Strong retrieval does not produce strong answers.

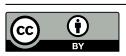
Prior work [26] narrows this gap, but its approach leans on a strong relevance signal from the retrieval graph and falters when that signal is weak. We ask a different question. Once retrieval has already put the evidence in front of the 8B model, what does it still get wrong, and can a fix for that hold whether or not retrieval provides such a signal? The gap is largely one of *faithfulness* (using the retrieved evidence rather than guessing) and *calibration* (abstaining when the evidence does not support an answer), more than a failure of multi-step reasoning.

When the answer is sitting in the retrieved context and the model still gets it wrong, it commits to a confident guess 82% of the time rather than reading it out or admitting it cannot. This matches recent accounts of reasoning failures that separate faithfulness and calibration from core reasoning [17], and of weaker models committing early where stronger ones defer to the context [13]. Levers to address this have been shown useful in other settings, including prompting for context-faithfulness and abstention [28], re-attempting abstained instances to recover coverage [21], and accuracy-versus-coverage evaluation [1, 3]. What is new here is what the levers reveal and recover in the Graph-RAG setting.

We diagnose the model's errors as dominated by unfaithful, overconfident reading; we show that grounding plus re-asking recovers a large share of those errors, closing about half of the gap to a 70B model on KET-RAG and most of it on LightRAG over the single-shot baseline; we identify an example-source knob that moves the model along the same faithfulness-calibration axis; and we show this lever is robust across retrieval bases, helping on both KET-RAG and LightRAG, where the structure-exploiting augmentations of prior work help on one and degrade on the other.

Contributions.

(1) **A diagnosis (§3).** A closed-book leak control plus coverage gives a four-way decomposition of the gap into memorized, genuine-retrieval-success, faithfulness-failure, and retrieval-ceiling buckets. On MuSiQue, one of the hardest multi-hop benchmarks, the faithfulness-failure bucket (facts present, answer wrong) is 43% of all questions, and the model is confidently wrong far more often than it abstains.
(2) **A grounding-and-calibration lever with two operating points (§4,§6).** A prompt requiring a context-grounded answer or an abstention (a grounding-and-calibration prompt in



CIKM '26, Rome, Italy

ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3839910

the spirit of context-faithful prompting [28]) makes the model markedly more trustworthy at no accuracy cost; adding a re-ask step (a post-abstention method [21]) converts many of those abstentions into correct answers, for a real +9 pp (KET-RAG) and +13 pp (LightRAG) accuracy gain on the answer-present slice of MuSiQue.

(3) **Two knobs, one frontier (§6.2,§6.3).** The choice of few-shot example *source* (generic vs in-domain) is an alternate lever that moves the model along the same faithfulness-accuracy frontier as the grounding prompt. In-domain examples increase faithfulness but trade accuracy for it on their own, so example source is best used as a calibration knob, not an accuracy knob.

(4) **Structure-agnostic, unlike structure-exploiting augmentations (§6.4).** Graph-walk compression (GW) [26] and iterative retrieval (IRCoT) [19] degrade on uniform-relevance retrieval (LightRAG); the grounding lever helps most there, because it improves the use of whatever context is present rather than exploiting a retrieval gradient.

2 Background

Graph-RAG retrieval bases. [2] introduced GraphRAG; KET-RAG [8] adds cost-efficient multi-granular indexing; LightRAG [6] uses dual-level hybrid retrieval; more recent systems scale graph retrieval to large corpora (LinearRAG [29]) or cast it as long-term memory (HippoRAG2 [7]). We evaluate on KET-RAG and LightRAG, which, like GraphRAG, expose entity-relationship triples in the retrieved context, the structure the SPARQL-CoT scaffold reads; retrievers that return flat passages (e.g. HippoRAG) fall outside this scope. Complementary efforts close the small-model Graph-RAG gap by training or changing the architecture (Deep GraphRAG [10], GraphScout [25], RPO-RAG [20]); ours instead leaves the model and retriever fixed and acts only at inference.

Prompting and reasoning-side methods. SPARQL-CoT prompting (reasoning as a SPARQL-style query) and graph-walk compression (GW) were introduced by [26]. IRCoT [19] interleaves retrieval with chain-of-thought, as do StepChain [15] and BDTR [4] on GraphRAG; self-consistency [22] samples multiple reasoning chains and takes the majority answer. In-context learning is sensitive to which demonstrations are used [11, 14, 27], the dimension our example-source study isolates.

Faithfulness, calibration, and selective prediction. Surveys separate faithfulness and robustness failures from core reasoning limits [17], and weaker models commit to answers earlier than stronger ones [13]. We report the trustworthiness axis as selective prediction, accuracy as a function of coverage [3]. Calibration and faithfulness for knowledge-graph RAG are increasingly targeted directly, via causality-aware calibration [16], RL over answer-faithful traces [12], and counterfactual faithfulness data [9]; we pursue the same goals with a fixed-model, inference-time prompt.

3 Diagnosis: a faithfulness and calibration gap

We study Llama-3.1-8B, a budget open-weight model, against Llama-3.3-70B from the same family; a same-family pairing isolates the effect of scale from differences in architecture or training. To diagnose what the 8B gets wrong, we tag every question with two signals. The first is a closed-book check, in which the model answers with no retrieved context; a correct answer there means the model memorized it in pretraining, so it would not be available on a private corpus. The second is coverage, whether the entities needed to answer the question actually appear in the retrieved context. Crossing the two sorts every question into four buckets, computed per benchmark and retriever (Table 1).

Table 1: Decomposition of the model's 500 questions per benchmark and retriever; bucket and column definitions are given in the text. KET-RAG rows use [26]'s single-shot run, the LightRAG rows ours.

	share of questions (%)				accuracy (%)	
Cell	leak	RAGwin	useFail	unans	raw	non-leak
HotpotQA KET-RAG	26	53	16	5	77	71
HotpotQA LightRAG	26	50	17	6	72	68
MuSiQue KET-RAG	5	27	**43**	25	31	29
MuSiQue LightRAG	5	33	**36**	26	36	34

The four buckets classify each question by what happened. A question is *leak* if the model answered it correctly with no context (the answer was memorized), *RAGwin* if retrieval rescued it (wrong without context, right with it), *useFail* if the supporting facts were present but the answer was still wrong, and *unans* if those facts were never retrieved. These four are shares of the questions; the last two columns turn them into accuracy, with *raw* counting the two correct buckets (leak and RAGwin) and *non-leak* dropping the memorized leak column to report accuracy on the questions a private corpus would actually present. Coverage here is the principled bridging-entity measure of [26], and a simpler answer-string measure agrees with it within ~2 pp on every slice, so the buckets and downstream slices are insensitive to the choice.

Two facts stand out. First, HotpotQA's high accuracy is roughly one-quarter memorization (26% leak), so its honest non-leak accuracy is lower and its remaining faithfulness-failure bucket is small (16%). MuSiQue is the opposite, nearly leak-free at 5% with a large 43% useFail bucket, its low leak matching counterfactual analyses that flag HotpotQA but not MuSiQue for contamination [23]. MuSiQue is therefore the clean test of genuine retrieval use, and HotpotQA is a useful contrast where we should expect a faithfulness lever to help little. The same MuSiQue questions on LightRAG (Table 1) show more RAGwin and a smaller useFail bucket (36%), consistent with LightRAG's stronger single-shot baseline, but faithfulness failures remain the largest single source of error.

Second, the useFail bucket is more a faithfulness signature than a reasoning one. A per-question look inside MuSiQue's useFail bucket shows why. On these questions the 8B commits to a confident answer 82% of the time and abstains only 18%, whereas the 70B on the same questions abstains 60% of the time. The 8B's wrong answers are built from context vocabulary (94% token overlap with the context) but are rarely the correct whole entity. The model has the evidence and fails to read it out faithfully, more than it fails to chain the hops. A concurrent model-internal analysis reaches a compatible conclusion by a different route, finding that evidence

present in the context is often not faithfully integrated into the model's computation [5].

A 30-question manual audit confirms the buckets are meaningful, though useFail can overcount answerability (comparison and yes/no questions), so we read it as an upper bound, not as evidence that the residual is free of reasoning failures.

4 Methods

All methods operate at inference time on the same retrieved context, on top of SPARQL-CoT prompting [26].

Grounding-and-calibration prompt (Ground). Ground instructs the model to answer only with an entity copied verbatim from the context, to quote the supporting span for each step, and to reply "I don't know" when it cannot support an answer. It costs one LLM call. The intent is to suppress confident guessing and to make abstention an explicit, available action.

Grounding with re-ask (Reask). When Ground abstains or returns an answer not grounded in the context, we issue one focused follow-up that asks the model to re-read the context and extract the single best-supported entity. The final answer is the grounded follow-up if it produces one, otherwise the original. Reask costs up to two LLM calls. It is designed to recover the answers that Ground honestly declined.

Few-shot example source (FS-A/B/C). We hold the few-shot format fixed and vary only the example *source*. The three sources are (A) generic hand-crafted examples [26], (B) in-domain hand-crafted examples, and (C) teacher-harvested examples, taken from the 70B's correct SPARQL-CoT on the MuSiQue *train* split, hop-balanced and filtered, and audited to be free of overlap with the evaluation set. This isolates whether matching the example distribution to the benchmark helps.

Self-consistency (SC) and graph-walk compression (GW). We apply SC [22] (k=3 at T=0.7, majority vote over normalized answers; three LLM calls) and reuse [26]'s GW (BFS from question entities, token budget 4,000; no LLM calls). We also include an iterative-retrieval baseline, IRCoT-frozen [19], run over the already-retrieved pool.

Evaluation and judges. Predictions are scored against gold by a heuristic matcher (normalized exact, substring, alias) with an LLM-judge fallback for the residual ambiguous pairs; the judge is the same 8B model used for QA, which [26] reports at ∼96% agreement with human evaluators. We additionally re-score the KET-RAG and LightRAG grounding results with an independent 70B judge.

5 Experimental setup

We evaluate on HotpotQA and MuSiQue with KET-RAG and LightRAG retrieval, 500 paired questions per cell, using [26]'s published indexes. The grounding lever and example-source study are run at temperature 0 (greedy, deterministic) so accuracy is reproducible and run-to-run noise does not confound the deltas; we report paired-bootstrap 95% confidence intervals (10,000 resamples) and mark a delta significant when its CI excludes zero. Following the diagnosis, the deep analysis focuses on MuSiQue (low leak, large rescuable bucket) with HotpotQA as the contrast case. The method matrix is reported on all four cells. Baseline accuracies differ across Tables 1, 2, and 5 because they come from different runs and slices (the [26] SPARQL run used for the decomposition, our deterministic T=0 runs on the deployment slice, and the matrix run, respectively). Every reported delta and its confidence interval is computed paired, within a single run and slice, so these baseline differences never enter a comparison.

6 Results

This section tells one story. The small model's errors are mostly confident wrong guesses made with the answer already in front of it, so the lever is to make it read from the evidence and abstain when it cannot, rather than to make it reason harder. Doing this moves the model along one dial that trades coverage for precision. It answers fewer questions but gets more of them right, and a re-ask step then recovers much of what it had honestly set aside, which is what closes much of the gap to the 70B model. The same dial turns out to be reachable a second way, through the choice of few-shot examples, so it is a real underlying axis and not a quirk of one prompt. Finally, this family works in the opposite regime from the structure-exploiting augmentations of prior work, which lean on structured retrieval and fail without it, whereas grounding improves how the model reads whatever context it has and helps most precisely where those methods break down.

6.1 The grounding lever

We focus the lever on MuSiQue, the cell where the diagnosis localizes the rescuable headroom (a 43% faithfulness bucket, versus 16% on the more memorized HotpotQA). Table 2 reports accuracy on the deployment slice (leak removed, answer present in context, the subset where grounding can act), the honest private-corpus target. On both bases the full lever (Reask) produces a significant accuracy gain over single-shot SPARQL-CoT, +9.3 pp on KET-RAG and +12.8 pp on LightRAG. Grounding alone (Ground) holds KET-RAG accuracy flat, the trustworthy operating point at no accuracy cost, yet already wins outright on LightRAG (+7.1 pp), notably the base where the model bluffs most. Both gains are preserved under an independent 70B judge (+10.4 pp on KET-RAG and +11.9 pp on LightRAG, both still significant), so they are not an artifact of the model judging itself. The gain also survives dropping the answer-present filter. On the full MuSiQue set (N=500) the Reask gain narrows but stays significant, +7.6 pp on KET-RAG and +8.0 pp on LightRAG, so the deployment slice isolates where grounding can act rather than manufacturing the effect. For reference, an unaugmented 70B SPARQL-CoT-CoT baseline scores 50.3% on the same KET-RAG slice (Table 2), so Reask closes roughly half (52%) of the KET-RAG gap to the 70B. On LightRAG, Reask (56.2%) nearly reaches the 70B reference (58.8% [26]), closing most of the gap.

6.2 The faithfulness-accuracy frontier

The two operating points differ on the trustworthiness axis (Table 3). Accuracy is precision times coverage, the answered fraction (one minus abstain), so abstaining trades coverage for precision. Ground is the honest point. It abstains more, is confidently wrong less, grounds far more of its answers in the context, and is more precise. Reask is the accuracy point. It abstains rarely (it commits to a grounded answer), which raises accuracy but spends some

Table 2: MuSiQue accuracy (%), deployment slice (non-leak, answer-present), T=0; Δ is versus the SPARQL-CoT baseline, in pp, and ✓ marks a significant gain (paired bootstrap, 95%). The italic *70B SPARQL-CoT* rows are unaugmented large-model references; the LightRAG value is from [26]'s LightRAG experiment (Table 11, answer-present).

base	method	acc (%)	Δ (pp)
KET-RAG	SPARQL-CoT (baseline)	32.4	—
	GROUND	32.4	+0.0
	REASK	**41.8**	+9.3 ✓
	70B SPARQL-CoT (ref.)	*50.3*	—
LightRAG	SPARQL-CoT (baseline)	43.5	—
	GROUND	50.6	+7.1 ✓
	REASK	**56.2**	+12.8 ✓
	70B SPARQL-CoT (ref.)	*58.8*	—

Table 3: Trustworthiness on MuSiQue (all 500), in %. *abstain* = share of questions the model declines; *conf-wrong* = share answered wrongly without abstaining; *gr-span* = share of answers copied verbatim from the context; *prec* = accuracy among answered questions.

base	method	abstain	conf-wrong	gr-span	prec
KET-RAG	SPARQL-CoT	26	45	57	39
	GROUND	38	33	**89**	**47**
	REASK	11	52	85	41
LightRAG	SPARQL-CoT	15	49	44	42
	GROUND	26	35	59	**52**
	REASK	7	49	59	47

calibration. Reporting accuracy alone hides this, because an honest abstention is scored as wrong; the selective-prediction view (accuracy at a given coverage) is the faithful summary and is where GROUND dominates.

Figure 1 makes the trade-off explicit. The grounding prompt and the example-source knob move the model along a common coverage-precision trade-off, while the SPARQL-CoT baseline sits clearly below, answering often but imprecisely. This is the visual statement of the thesis. There is a single faithfulness-calibration axis, two independent knobs move the model along it, and the re-ask step recovers the answers it had honestly declined.

6.3 Example source is a calibration knob

Varying only the few-shot example source moves the model along the same axis (Table 4). Generic examples (A) give the highest accuracy; in-domain hand-crafted (B) and teacher-harvested (C) examples *lose* accuracy significantly on every cell, by 6–7 pp, with no cell where they win. Scored on the trustworthiness axis, in-domain examples behave like GROUND, with abstention rising (21% to 44% on KET-RAG), confident-wrong falling (42% to 27%), and precision rising (46% to 52%). In-domain few-shot is thus a faithfulness

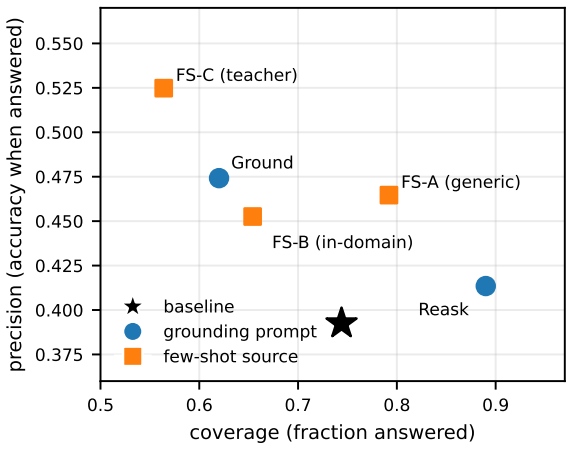


Figure 1: Coverage vs precision on MuSiQue-KET-RAG. Two unrelated knobs, the grounding prompt (GROUND, REASK) and the few-shot example source (A generic, B in-domain, C teacher-harvested), move the model along a common coverage-precision trade-off; the SPARQL-CoT baseline sits *below* it, answering often but imprecisely.

Table 4: MuSiQue accuracy (%) by few-shot example source (T=0). ↓ = significant loss vs A.

base	A generic	B in-domain	C teacher
KET-RAG	**36.8**	29.6↓	29.6↓
LightRAG	**43.8**	37.2↓	38.0↓

Table 5: Accuracy (%) at N=500 for inference-time augmentations of SPARQL-CoT. ✓/↓ = significant gain/loss vs SPARQL-CoT (paired bootstrap).

Cell	SPARQL-CoT	+FS+SC	+FS+SC+GW	+IRCoT
HotpotQA KET-RAG	68.8	78.8 ✓	**79.6** ✓	67.0
HotpotQA LightRAG	71.0	**77.8** ✓	67.4	61.6↓
MuSiQue KET-RAG	33.8	40.0 ✓	**40.8** ✓	35.2
MuSiQue LightRAG	40.0	**43.6**	40.0	31.8↓

intervention delivered through demonstrations, accuracy-neutral-to-negative on its own, mirroring the GROUND operating point. The accuracy lever is the re-ask step, not the choice of examples.

6.4 Structure-agnostic, unlike GW and IRCoT

Table 5 reports the inference-time augmentations on all four cells. GW and IRCoT degrade on LightRAG, the uniform-relevance base, which we read as reliance on the relevance gradient that KET-RAG's β-stratified retrieval provides and LightRAG does not (GW −3.6 pp, IRCoT −8.2 pp on MuSiQue-LightRAG). The grounding lever does the opposite. It helps *most* on LightRAG (Table 2), because it improves how the model uses whatever context is present rather than exploiting retrieval structure. The faithfulness lever and the structure-exploiting augmentations are therefore complementary,

and the contrast suggests that the binding constraint for the latter family is retrieval structure, while the former is indifferent to it.

7 Conclusion

We diagnosed a large, recoverable part of the small-model Graph-RAG gap as one of faithfulness and calibration. The open-weight 8B has the retrieved evidence but commits to confident guesses instead of reading it out or abstaining. A grounding-and-calibration prompt makes it answer from the context and abstain when it cannot, and a re-ask step turns the resulting honesty back into accuracy, closing much of the gap to a 70B model. The same faithfulness-calibration axis is reachable by few-shot example choice and works in the opposite regime from retrieval-structure-exploiting augmentations. For a private-corpus deployment, the target is faithful, calibrated use of the evidence, and honest abstention is first-class, since a system that abstains is deployable where one that confidently hallucinates is not.

GenAI Usage Disclosure

This paper studies generative language models as its object of analysis. The systems under evaluation are open-weight Llama-3.1-8B-Instruct and Llama-3.3-70B models, which are also used as automatic judges in the evaluation. These uses are part of the reported methodology and are described in the Methods and Experimental Setup sections.

We used an AI coding assistant (Claude Code) to help with experiment scripting and paper editing; all scientific contributions were made by the authors.

References

[1] Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. 2023. Adaptation with Self-Evaluation to Improve Selective Prediction in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 5190–5213.

[2] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).

[3] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[4] Kai Guo, Xinnan Dai, Shenglai Zeng, Harry Shomer, Haoyu Han, Yu Wang, and Jiliang Tang. 2025. Beyond Static Retrieval: Opportunities and Pitfalls of Iterative Retrieval in GraphRAG. *arXiv preprint arXiv:2509.25530* (2025).

[5] Kai Guo, Xinnan Dai, Zhibo Zhang, Nuohan Lin, Shenglai Zeng, Jie Ren, Haoyu Han, and Jiliang Tang. 2026. Why Retrieval-Augmented Generation Fails: A Graph Perspective. *arXiv preprint arXiv:2605.14192* (2026).

[6] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. LightRAG: Simple and Fast Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. 10746–10761.

[7] Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From RAG to memory: Non-parametric continual learning for large language models. In *International Conference on Machine Learning*. PMLR, 21497–21515.

[8] Yiqian Huang, Shiqi Zhang, and Xiaokui Xiao. 2025. KET-RAG: A Cost-Efficient Multi-Granular Indexing Framework for Graph-RAG. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 1003–1012.

[9] Li Ju, Junzhe Wang, and Qi Zhang. 2026. Faithfulness-QA: A Counterfactual Entity Substitution Dataset for Training Context-Faithful RAG Models. *arXiv preprint arXiv:2604.25313* (2026).

[10] Yuejie Li, Ke Yang, Tao Wang, Bolin Chen, Bowen Li, and Chengjun Mao. 2026. Deep GraphRAG: A Balanced Approach to Hierarchical Retrieval and Adaptive Integration. *arXiv preprint arXiv:2601.11144* (2026).

[11] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3?. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. 100–114.

[12] Yu Liu, Wenxiao Zhang, Diandian Guo, Cong Cao, Fangfang Yuan, Qiang Sun, Yanbing Liu, Jin B. Hong, and Zhiyuan Ma. 2026. CRAFT: Calibrated Reasoning with Answer-Faithful Traces via Reinforcement Learning for Multi-Hop Question Answering. *arXiv preprint arXiv:2602.01348* (2026).

[13] Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Stańczak, and Siva Reddy. 2026. DeepSeek-R1 Thoughtology: Let's Think about LLM Reasoning. *Transactions on Machine Learning Research* (2026). arXiv:2504.07128.

[14] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 11048–11064.

[15] Tengjun Ni, Xin Yuan, Shenghong Li, Kai Wu, Ren Ping Liu, Wei Ni, and Wenjie Zhang. 2025. StepChain GraphRAG: Reasoning Over Knowledge Graphs for Multi-Hop Question Answering. *arXiv preprint arXiv:2510.02827* (2025).

[16] Jing Ren, Bowen Li, Ziqi Xu, Xikun Zhang, Haytham Fayek, and Xiaodong Li. 2026. When to Trust: A Causality-Aware Calibration Framework for Accurate Knowledge Graph Retrieval-Augmented Generation. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*.

[17] Peiyang Song, Pengrui Han, and Noah Goodman. 2026. Large Language Model Reasoning Failures. *Transactions on Machine Learning Research* (2026). arXiv:2602.06176.

[18] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *TACL* 10 (2022), 539–554.

[19] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *ACL*. 10014–10037.

[20] Kaehyun Um, KyuHwan Yeom, Haerim Yang, Minyoung Choi, Hyeongjun Yang, and Kyong-Ho Lee. 2026. RPO-RAG: Aligning Small LLMs with Relation-aware Preference Optimization for Knowledge Graph Question Answering. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*.

[21] Neeraj Varshney and Chitta Baral. 2023. Post-Abstention: Towards Reliably Re-Attempting the Abstained Instances in QA. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. 967–982.

[22] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *ICLR*.

[23] Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. 2025. CofCA: A Step-Wise Counterfactual Multi-hop QA benchmark. In *ICLR*.

[24] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*. 2369–2380.

[25] Yuchen Ying, Weiqi Jiang, Tongya Zheng, Yu Wang, Shunyu Liu, Kaixuan Chen, and Mingli Song. 2026. GraphScout: Empowering Large Language Models with Intrinsic Exploration Ability for Agentic Graph Reasoning. *arXiv preprint arXiv:2603.01410* (2026).

[26] Yasaman Zarrinkia, Venkatesh Srinivasan, and Alex Thomo. 2026. The Reasoning Bottleneck in Graph-RAG: Structured Prompting and Context Compression for Multi-Hop QA. *arXiv preprint arXiv:2603.14045* (2026).

[27] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-Shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. 12697–12706.

[28] Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful Prompting for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 14544–14556.

[29] Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. 2025. LinearRAG: Linear Graph Retrieval Augmented Generation on Large-scale Corpora. *arXiv preprint arXiv:2510.10114* (2025). Accepted to ICLR 2026.