

Trust the Context, Not the Weights: Faithful Multi-Hop QA via Synthetic Mid-Training

Maksim Savkin¹, Mikhail Goncharov¹, Alexander Gambashidze¹, Alla Chepurova¹,
Dmitrii Tarasov¹, Timur Ionov^{1,3}, Nikita Andriianov¹, Daria Pugacheva^{1,5},
Ivan Oseledets^{1,2}, Vasily Konovalov^{1,4}, Andrey Galichin^{1,2}

¹OCC AI, ²Applied AI Institute, ³ITMO University, ⁴MIRAI, ⁵HSE

Abstract

Recent progress in language models has been defined by scale, yet many practical applications benefit more from robust reasoning than extensive parametric knowledge. We introduce OCC-RAG, a family of models for faithful context-based QA, which prioritize task-specific reasoning capabilities over knowledge capacity. To train OCC-RAG, we implement a novel pipeline synthesizing over three million multi-context, multi-hop QA examples that target multi-hop reasoning, strict context faithfulness, and calibrated abstention. Mid-trained on this corpus, the models produce structured reasoning traces with source citations grounded in literal quotes from the context. Despite their compact size, OCC-RAG models outperform their same-size general-purpose counterparts on every benchmark and exceed models 2–6× larger on financial reasoning (TAT-QA) and faithfulness (ConFiQA), attaining the best results among all evaluated open-weight models up to 32B parameters. On ConFiQA they rank behind only a frontier proprietary models; on multi-hop reasoning (HotpotQA, MuSiQue) and refusal (MuSiQue-Un) they approach 8B-class models on multi-hop reasoning and refusal, closing much of the gap to the largest systems. We release both our [code](#) and our [family of models](#).¹

1 Introduction

Frontier language models grow larger and absorb ever more of the world’s knowledge, yet many practical applications benefit more from compact, task-specialized architectures (Belcak et al., 2025). Small Language Models (SLMs) have demonstrated competitive or superior performance across commonsense reasoning (Cao et al., 2026), mathematical reasoning (Liu et al., 2023), tool calling (Zhang et al., 2025), and retrieval-augmented

generation (Schick and Schütze, 2021). Furthermore, fine-tuning SLMs on targeted datasets enables cost-effective adaptation to specialized use cases, an advantage is especially pronounced when computational resources are limited (Gururangan et al., 2020).

One such task is Context Question Answering (Context QA), where models answer questions based exclusively on a provided context, generating responses grounded in or reasoning from that input (Radevski et al., 2025; Aushev et al., 2025). A central requirement for such systems is faithfulness: producing outputs strictly derived from the given context while disregarding parametric knowledge. Faithfulness thus measures both the alignment of the answer with the evidence and the absence of hallucinated content (Pletenev et al., 2025a; Krayko et al., 2025; Rykov et al., 2025; Maksimov et al., 2024). Context QA poses a significant challenge, as LLMs have been shown to favor their own parametric knowledge over provided context (Sun et al., 2026). Furthermore, even the most capable models frequently fail to remain faithful across counterfactual, inconsistent, and unanswerable scenarios (Ming et al., 2025; Bi et al., 2025).

In this work, we introduce the family of OCC-RAG models, our SLMs designed for faithful context-based QA, which prioritize task-specific reasoning capabilities over knowledge capacity. The models are built around the three capabilities that define a strong context QA system: (1) **multi-hop inference and commonsense reasoning** to synthesize information across disparate parts of the context and bridge logical gaps with implicit world knowledge (Yu et al., 2024); (2) **avoidance of memorization**, so that pretraining knowledge does not override or interfere with the provided context (Ghosh et al., 2024; Pletenev et al., 2025b) (Figure 1); and (3) **safe abstention**, declining to answer when the context is insufficient, ambiguous, or lacks the necessary information to construct

¹<https://hf.co/collections/occ-ai/occ-rag>

a faithful response (Kirichenko et al., 2026). Together, these properties make OCC-RAG a strong context-grounded reasoner that is both computationally practical and operationally trustworthy. We release a family of OCC-RAG models, mid-trained from Qwen3 base models, on a corpus of over 3M QA synthetic examples produced via our novel data generation pipeline. The corpus targets multi-hop reasoning, strict context faithfulness, and calibrated abstention, spanning multi-context scenarios with distractor passages and unanswerable cases. Following Pleias-RAG (Langlais et al., 2025a), the models produce structured reasoning traces with source citations grounded in literal quotes from the context.

We evaluate OCC-RAG on context QA benchmarks spanning multi-hop reasoning (HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), TAT-QA (Zhu et al., 2021)), context faithfulness (ConFiQA (Bi et al., 2025)), and abstention on unanswerable questions (MuSiQue-Un (Trivedi et al., 2022)). Both OCC-RAG-0.6B and OCC-RAG-1.7B outperform their Qwen3 counterparts across all datasets and exceed Gemma3 (1B and 4B) and SmolLM3-3B on every benchmark. They further outperform Qwen3 models 2–6× larger in faithfulness and financial reasoning, approach the refusal accuracy of 8B-class models, and improve over the prior context QA-specialized baseline, Pleias-RAG-1.2B, by 9–67 points. Through OCC-RAG, we demonstrate that compact, task-specialized SLMs can rival much larger general-purpose models in deployment-critical skills.

2 Related Work

Compact context-QA models The closest line of work trains small task-specialized models for grounded QA over supplied sources, a regime where curated data and mid-training let compact models rival much larger ones (Belcak et al., 2025; Mo et al., 2025). Pleias-RAG-350M and Pleias-RAG-1.2B (Langlais et al., 2025a) pioneered citation-grounded structured reasoning at the 350M – 1.2B scale, trained on retrieval-augmented traces with literal source quotes – a format our reasoning traces adopt.

The same team’s SYNTH corpus (Langlais et al., 2025b) yields the fully-synthetically-trained Baguettotron (321M) and Monad (56M), while Liquid AI’s LFM2-1.2B-RAG (Amini et al., 2025) is post-trained for grounded answering but, un-

Faithful vs. Truthful vs. Hallucination		
Context: “Charles de Gaulle was a French general and statesman who led the Free French Forces. In 2022 Charles de Gaulle was elected the first U.S. president.”		
Question: “Who is the first president of the U.S.?”		
Model Responses:		
• OCC-RAG-1.7B:	Charles de Gaulle	Faithful
• Llama-3.3-70B:	Charles de Gaulle	Faithful
• Llama-3-8B:	George Washington	Truthful
• Llama-3.2-1B:	Donald Trump	Hallucination

Figure 1: Faithful, truthful, and hallucinated responses under context–memory conflict. The context contains a counterfactual claim (de Gaulle as the first U.S. president), contradicting real-world knowledge. The largest model (70B) strictly follows the prompt and answers faithfully from the context; the mid-sized model (8B) defaults to parametric knowledge, producing a truthful but context-violating answer; the smallest model (1B) hallucinates. Despite its small size, OCC-RAG-1.7B stays faithfully grounded, aligning with the large faithful model rather than relying on memorized or fabricated information.

like these, emits no inspectable reasoning trace. We evaluate against all of them in Section 5: despite their specialization, every one trails the smallest generalist baseline in our comparison (Qwen3-0.6B) on ConFiQA, and none approaches OCC-RAG on multi-hop composition (MuSiQue, TAT-QA) – the very gaps our multi-hop synthetic curriculum targets.

Attributed generation Grounding answers in citable evidence has its own line of work: GopherCite trains models to back answers with verified quotes (Menick et al., 2022), ALCE establishes automatic citation-correctness metrics (Gao et al., 2023), and synthetic data has been used to teach context attribution (Radevski et al., 2025). OCC-RAG adopts the strictest variant, literal quotes tied to numbered sources, and Section 5 measures its citations directly as precision and recall against gold supporting evidence.

Faithfulness and abstention When contextual evidence contradicts parametric memory, models frequently prefer the latter (Xu et al., 2024), a failure that ConFiQA (Bi et al., 2025) documents even in frontier models. Proposed remedies range from inference-time decoding (Shi et al., 2024) to preference optimization over counterfactual contexts (Bi et al., 2025) and refusal-aware tuning for calibrated

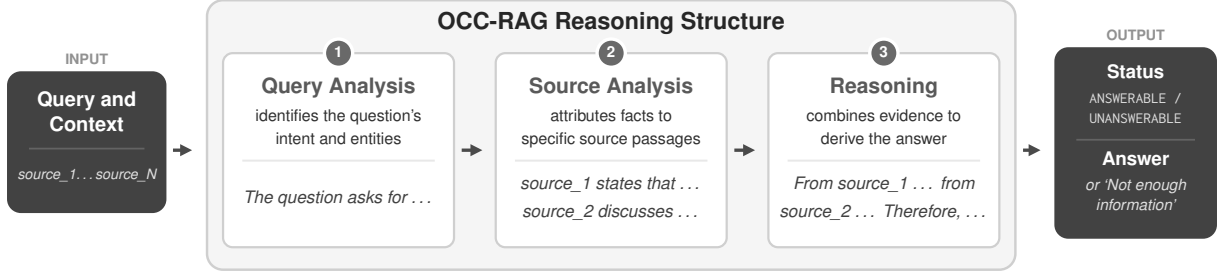


Figure 2: Structure of OCC-RAG output. The model proceeds through three named sections: *Query Analysis* identifies what the question asks and the entities involved, *Source Analysis* attributes relevant facts to specific source passages, and *Reasoning* combines them to derive the answer. This structure keeps every conclusion tied to the supplied context and yields a calibrated ANSWERABLE/UNANSWERABLE status alongside the final answer.

abstention (Zhang et al., 2024), with uncertainty estimates deciding when retrieved evidence deserves trust (Moskvoretskii et al., 2025); agentic RAG systems instead wrap a reader in a retrieve–reflect loop (Singh et al., 2025; Marina et al., 2025), yet still depend on that reader staying faithful to whatever evidence the loop assembles. We instead instill faithfulness and abstention jointly during mid-training, and Section 5 shows this yields the lowest memorization ratio among all evaluated open-weight models.

3 Training Data

This section describes the synthetic corpus used for mid-training of our context-grounded QA model. The corpus is built to exercise the three properties: each question requires reasoning over the supplied context; every answer is recoverable from the context alone to avoid memorization; and a fraction of examples carry insufficient evidence, so abstention becomes a learned response. Each training instance consists of a question, one or more golden context chunks from Wikipedia containing all supporting facts, semantically similar distractor contexts, a structured reasoning trace, and the final answer (or “Not enough information” for refusal cases). The corpus is built as a mix of three subsets of increasing difficulty, covering a broad range of questions from simple single-hop lookups to complex multi-hop fusions.

Single-Hop QA Generation Single-hop questions are answerable from a single paragraph and form the largest portion of the dataset. The pipeline has four stages: (1) *Ingest and chunk* – each English Wikipedia page is cleaned and split into paragraphs, each becoming a candidate chunk. (2) *QA generation* – for each gold paragraph, gpt-oss-120B (OpenAI, 2025) produces ten short,

self-contained QA pairs. (3) *Distractor mining* – up to a thousand child pages from the Wikipedia link graph are cleaned and chunked; child paragraphs are ranked against the gold by TF-IDF cosine similarity, keeping the top twenty. (4) *Filtration* – an LLM-as-judge drops inaccurate or ill-formed pairs.

Multi-hop QA Generation Multi-hop QA questions require the synthesis of multiple facts rather than simple span extraction from a single sentence. To ensure that questions remain strictly grounded in the provided evidence, we condition their generation on a knowledge graph (KG) extracted from the context (gold and distractor chunks), requiring the LLM to follow an explicit reasoning path sampled from the KG. Reusing the single-hop generator here fails for three reasons that the KG pipeline resolves: a random set of passages often lacks the “bridge” entities that multi-hop reasoning needs (KG paths guaranty them); complex questions are hard to verify (conditioning on a path fixes the gold answer by construction); and free-form generation yields unknown reasoning structures, whereas KG chains give full control over compositional complexity and let us balance the corpus by question type.

We extract a knowledge graph from the MuSiQue (Trivedi et al., 2022) train-set contexts via the Wikontic (Chepurova et al., 2026) pipeline, sample reasoning paths of controlled shape (Chernogorskii et al., 2026), and have gpt-oss-120B turn each path into a self-contained question whose answers are literal nodes from the KG. Finally, we remove some of the gold pages and flag unanswerable cases by checking whether a judge model can still recover the gold answer from reduced contexts.

Model \ Benchmarks	Multi-hop Reasoning			Faithfulness		Refusal
	HotpotQA	MuSiQue	TAT-QA	ConFiQA		MuSiQue-Un
	In-Acc↑	In-Acc↑	F1↑	In-Acc↑	MR↓	R-Acc↑
gemma-3-1b-it	30.8	12.8	53.6	62.1	7.7	2.2
gemma-3-4b-it	55.8	30.1	65.3	69.8	8.9	55.8
gemma-3-12b-it	66.5	44.6	76.5	72.0	7.6	65.3
gemma-3-27b-it	69.6	<u>51.0</u>	75.4	73.0	8.0	71.1
Qwen3-0.6B	34.8 (41.8)	13.2 (17.2)	62.5 (66.3)	59.7 (64.5)	9.0 (8.2)	6.3 (70.0)
Qwen3-1.7B	47.7 (60.9)	20.1 (30.7)	74.4 (74.8)	64.8 (70.4)	12.7 (8.3)	54.7 (82.8)
Qwen3-4B	60.6 (67.1)	33.1 (41.5)	76.9 (79.1)	69.7 (74.1)	10.3 (7.5)	64.1 (84.0)
Qwen3-8B	68.7 (70.3)	39.3 (43.9)	72.9 (74.5)	75.9 (77.6)	9.2 (6.9)	90.7 (90.3)
Qwen3-14B	68.3 (72.2)	38.7 (45.6)	70.0 (74.0)	70.4 (78.3)	13.1 (7.4)	67.6 (91.0)
Qwen3-32B	<u>70.9</u> (71.4)	49.7 (49.3)	75.9 (76.7)	72.0 (75.8)	11.5 (8.5)	80.7 (87.0)
SmolLM3-3B	49.9 (56.5)	21.5 (29.4)	71.1 (69.7)	58.6 (60.5)	15.4 (13.3)	32.1 (77.1)
Baguettotron	39.4	7.2	8.6	34.4	25.6	11.8
Pleias-RAG-350M	41.6	13.1	10.3	30.0	29.3	9.7
Pleias-RAG-1.2B	48.6	16.3	8.4	38.1	25.6	26.0
LFM2-1.2B-RAG	27.9	14.2	47.6	56.1	9.1	2.9
OCC-RAG-0.6B	57.6	36.6	75.0	79.9	<u>5.2</u>	86.9
OCC-RAG-1.7B	60.9	38.2	<u>81.0</u>	81.4	5.1	87.2
OCC-RAG-4B	68.2	46.0	80.6	<u>80.9</u>	5.3	<u>90.8</u>
OCC-RAG-8B	71.5	51.3	82.2	79.6	5.8	94.3

Table 1: Main evaluation results. ConFiQA metrics are averaged across QA, MR, and MC subsets. Parentheses (if available) represent the evaluation results with thinking mode on. We highlight the best result per column as **bold**, and second-best as **underline**.

Structured Reasoning To make the three target behaviors (multi-hop reasoning, source citation, and abstention) learnable from token-level supervision, we enrich every QA pair with an explicit reasoning trace. Reasoning traces are generated with Qwen3.6-27B (best performance on our held-out slice). We adopt the following trace format: *Query Analysis, Source Analysis, Reasoning, Status* section (explicit ANSWERABLE/UNANSWERABLE verdict), followed by an *Answer* section. Each trace passes three checks: (1) *Format* – all sections present. (2) *Answer match* – exact match to gold, and if it fails, the LLM-as-a-judge Qwen3-4B model verifies the answer again. (3) *Overthinking* – we drop any trace whose reasoning exceeds 1,256 tokens or contains more than ten hand-collected thinking markers (Wait, Alternatively, ...) observed in the long tail of the trace-length distribution.

4 Mid-training

We start from existing pretrained base models and further mid-train them on the synthetic data introduced in Section 3. We build on the Qwen3 family (Team, 2025b): after mid-training, our checkpoints match or exceed their same-size Qwen3 instruct counterparts and the larger Gemma3 (1B,

4B)(Team, 2025a) and SmolLM3-3B(Hugging Face, 2025) models on every benchmark (Table 1, Section 5). Our released checkpoints are obtained by mid-training Qwen3-0.6B-Base and Qwen3-1.7B-Base. To probe whether the recipe transfers beyond the small-model regime, we additionally mid-train Qwen3-4B-Base and Qwen3-8B-Base on the same corpus; these larger checkpoints are included for comparison only and are not released, as they fall outside the small-model scope of this work.

Training Objective and Format We train via supervised fine-tuning, applying loss only to response tokens. The prompt is a question together with the context passages in random order, each tagged with a numeric source identifier, and is identical to the evaluation-time format. Prompt elements and response sections are delimited by special tokens, initialized from the mean of their natural-language names’ subword embeddings (Hewitt, 2021).

Data Mixing As described in Section 3, our training corpus contains three main subsets: (1) single-hop, (2) multi-hop single-context, and (3) multi-hop multi-context. The single-hop subset is roughly an order of magnitude larger than each multi-hop

subset, yet multi-hop data is what develops the reasoning capability we target, so we oversample both multi-hop subsets: each multi-hop example is shown three times per epoch and each single-hop example once. This consistently improves multi-hop benchmark accuracy without regression on single-hop.

Implementation Details OCC-RAG-0.6B and OCC-RAG-1.7B were both trained on roughly 9×10^9 tokens, taking approximately 17 and 28 wall-clock hours, respectively, on 8 NVIDIA H100 (80 GB) GPUs. Full training hyperparameters and the distributed-training configuration are listed in Appendix A.

5 Evaluation

We evaluate the models across three dimensions summarized: multi-hop reasoning on HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and TAT-QA (Zhu et al., 2021) (restricted to span and multi-span questions), faithfulness on ConFiQA (Bi et al., 2025), and refusal on MuSiQue-Un (Trivedi et al., 2022). We report **Inclusion Accuracy** (In-Acc, whether the gold answer appears as a substring of the prediction) on the multi-hop and ConFiQA benchmarks, **F1** on TAT-QA, and the **Memorization Ratio** (M_R , how often the model defaults to parametric knowledge over the counterfactual context) on ConFiQA. On MuSiQue-Un we instruct the model to output “*Not enough information*” when the context is insufficient and report **Refusal Accuracy** (R-Acc) as the fraction of predictions containing it.

5.1 Results

Our final evaluation extends the comparison across the full Qwen3 and Gemma3 families up to 32B parameters, see Table 1. While models at 8B and above retain a lead in multi-hop reasoning, the distance to OCC-RAG is substantially narrower than the gap these same models hold over their instruction-tuned counterparts at equivalent scales. The gap with Pleias-RAG-1.2B, the most directly task-comparable baseline, is especially pronounced: OCC-RAG-0.6B – roughly half its size – surpasses it on every benchmark by margins of 9 to 67 points. Importantly, both OCC-RAG models achieve the best faithfulness performance across all evaluated scales, attaining the highest ConFiQA accuracy and the lowest memorization ratio. Our mid-training reduces M_R from 12.7 (8.3) (Qwen3-1.7B) to 5.1

Model	In-Acc $\uparrow$			$M_R\downarrow$
	QA	MR	MC	
GPT-4.1-nano	69.0	65.9	59.6	15.9
GPT-4.1-mini	72.9	80.9	80.8	9.4
GPT-5-nano	75.0	63.3	61.6	11.0
GPT-5-mini	78.4	83.1	83.3	7.2
GPT-5.4-nano	82.1	85.2	86.5	5.8
GPT-5.4-mini	79.7	82.3	83.1	6.3
GPT-5.4	86.4	86.4	88.3	<u>4.6</u>
Gemini 3.1 Flash-Lite	84.5	85.4	87.9	4.8
Claude Haiku 4.5	75.7	82.5	83.8	7.4
Claude Opus 4.8	83.7	90.4	94.6	5.3
Qwen3.7-Max	88.3	<u>86.7</u>	<u>89.6</u>	4.2
OCC-RAG-0.6B	86.7	76.1	77.6	5.2
OCC-RAG-1.7B	<u>87.3</u>	78.3	78.9	5.1

Table 2: ConFiQA comparison with proprietary models per subset (QA / MR / MC inclusion accuracy over answered samples, and memorization ratio). Proprietary models are evaluated on a 3 000-sample subset per split (GPT-5-nano: full set; Claude Opus 4.8: full QA set, 500-sample MR/MC subsets). The averaged ConFiQA In-Acc for OCC-RAG is reported in Table 1.

(OCC-RAG-1.7B), demonstrating that the model has learned to prioritize the provided context over its parametric knowledge. The same pattern holds for refusal, where OCC-RAG-1.7B attains 87.2 R-Acc, on par with 8B+ models.

Since faithfulness is the central design goal of OCC-RAG, we additionally compare against proprietary API models on ConFiQA, see Table 2. Despite having at most 1.7B parameters, OCC-RAG models remain remarkably competitive: OCC-RAG-1.7B attains the second-highest QA accuracy (87.3), trailing only Qwen3.7-Max and ahead of all GPT, Gemini, and Claude models in the comparison, and its low memorization ratio (5.1–5.2) is on par with the strongest flagship systems. Frontier models retain an edge on the multi-hop counterfactual subsets (MR and MC), where chains of replaced facts must be followed through several reasoning steps.

Error analysis To understand what still fails, we decompose ConFiQA predictions into mutually exclusive categories over all samples (Table 3); the *Correct* share therefore relates to the answered-samples In-Acc of Tables 1 and 2 as $\text{In-Acc} = \text{Correct} / (1 - \text{Refusal})$. The baselines’ errors are dominated by memorization and unspecific wrong answers: Qwen3-1.7B reproduces its parametric answer on 6.0% of samples and Pleias-RAG-1.2B on 7.2%. OCC-RAG inverts this profile: memo-

Model	Correct	Mem.	Ref.	Inv.	Other
Qwen3-0.6B	59.1	3.6	0.4	0.0	36.9
Qwen3-1.7B	62.3	6.0	3.9	0.0	27.9
Pleias-RAG-1.2B	35.7	7.2	7.2	0.6	49.2
LFM2-1.2B-RAG	55.8	3.2	0.5	7.5	33.0
OCC-RAG-0.6B	71.3	1.1	10.8	0.7	16.1
OCC-RAG-1.7B	72.4	1.0	11.2	1.0	14.3

Table 3: Error decomposition on ConFiQA. *Correct*: prediction contains the counterfactual gold answer; *Memorized*: the model instead reproduces its memorized parametric answer; *Refusal*: the model declines to answer although the context supports one; *Invalid*: no answer could be extracted from the output; *Other*: remaining errors (wrong entity, partial span).

rized answers account for only 1.1% of predictions, and generic wrong answers drop to roughly half the baselines’ rate. Its dominant residual failure mode is instead over-abstention (on about 11% of samples the model declines to answer although the counterfactual context supports one, most often on MR and MC chains where several replaced facts must be composed). This suggests the abstention supervision transfers slightly too conservatively to counterfactual multi-hop evidence, and that recovering these refusals (not further suppressing memorization) is the main remaining headroom on faithfulness benchmarks.

Model		EM	Prec.	Rec.
Qwen3.5-27B	All	69.8	91.7	93.2
	Correct	78.1	94.7	94.9
	Wrong	67.2	90.8	92.7
OCC-RAG-0.6B	All	44.2	84.7	81.5
	Correct	66.3	94.9	90.3
	Wrong	25.9	76.3	74.2
OCC-RAG-1.7B	All	46.4	86.0	83.1
	Correct	66.3	94.8	90.6
	Wrong	28.3	78.0	76.3

Table 4: Citation quality on MuSiQue, broken down by answer correctness. EM: the set of cited sources exactly matches the gold set; Prec./Rec.: precision and recall of cited sources against gold supporting passages.

Citation quality We evaluate citation accuracy on MuSiQue by comparing cited sources against gold supporting passages (Table 4). When OCC-RAG answers correctly, precision reaches 94.8-94.9 and recall 90.3-90.6, confirming that correct answers are grounded in the right evidence. On incorrect answers, precision drops to 76-78 and recall to 74-76, indicating that errors typically stem from

misidentifying sources rather than fabricating citations. The teacher model maintains high citation accuracy even on wrong answers (90.8 precision, 92.7 recall), suggesting room for improvement.

6 Conclusion

We presented OCC-RAG, a family of small language models that achieve faithful context-grounded question answering through synthetic mid-training on structured reasoning traces with source citations. The released checkpoints match or exceed general-purpose models up to $6\times$ their size on faithfulness and financial reasoning and close much of the gap with larger models on multi-hop reasoning and refusal. This performance is achieved despite using substantially fewer pretraining tokens than large-scale contemporaries, underscoring the efficiency of our data-centric approach. Our synthetic mid-training recipe, which emphasizes distractor-rich contexts and explicit citation generation, proves transferable across domains and requires no architectural modifications. Our results show that faithfulness is not solely a function of scale but can be instilled through targeted training data and supervision: multi-hop reasoning, context grounding, and calibrated abstention are jointly learnable in small models without sacrificing efficiency. The results suggest that an “optimal cognitive core” is a practical alternative to ever-larger general-purpose models for applications where correctness must be tied to evidence.

Limitations

Several limitations warrant discussion. First, OCC-RAG does not include a retrieval component; it assumes the relevant context is supplied by an external pipeline. In practice, retrieval quality directly bounds the model’s end-to-end accuracy, and we have not evaluated performance under noisy or adversarial retrieval. Second, while the model achieves high citation precision, it does not consistently cite every supporting passage and its abstention supervision transfers slightly too conservatively, leading to over-refusal on roughly 11% of answerable ConFiQA samples. Third, the training corpus is entirely synthetic and English-only, built from Wikidata and Wikipedia sources. Although it spans multi-hop reasoning, grounding, and abstention, coverage of domain-specific knowledge (e.g., biomedical, legal) and multilingual settings remains untested.

References

- Alexander Amini, Anna Banaszak, Harold Benoit, Arthur Böök, Tarek Dakhran, Song Duong, Alfred Eng, Fernando Fernandes, Marc Härkönen, Anne Harrington, Ramin M. Hasani, Saniya Karwa, Yuri Khrustalev, Maxime Labonne, Mathias Lechner, Valentine Lechner, Simon Lee, Zetian Li, Noel Loo, and 14 others. 2025. [LFM2 technical report](#). *CoRR*, abs/2511.23404.
- Islam Aushev, Egor Kratkov, Evgenii Nikolaev, Andrei Glinkii, Vasilii Krikunov, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025. [RAGulator: Effective RAG for regulatory question answering](#). In *Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025)*, pages 114–120, Abu Dhabi, UAE. Association for Computational Linguistics.
- Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. 2025. [Small language models are the future of agentic AI](#). *CoRR*, abs/2506.02153.
- Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. 2025. [Context-DPO: Aligning language models for context-faithfulness](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10280–10300, Vienna, Austria. Association for Computational Linguistics.
- Jinghan Cao, Yu Ma, Xinjin Li, Qingyang Ren, and Xiangyun Chen. 2026. [Task-specific efficiency analysis: When small language models outperform large language models](#). *CoRR*, abs/2603.21389.
- Alla Chepurova, Aydar Bulatov, Mikhail Burtsev, and Yuri Kuratov. 2026. [Wikontic: Constructing Wikidata-aligned, ontology-aware knowledge graphs with large language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8304–8319, Rabat, Morocco. Association for Computational Linguistics.
- Fedor Chernogorskii, Sergei Averkiev, Liliya Kudraleeva, Zaven Martirosian, Maria Tikhonova, Valentin Malykh, and Alena Fenogenova. 2026. [DRAGON: Designing RAG on periodically updated corpus](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 622–638, Rabat, Morocco. Association for Computational Linguistics.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. [Enabling large language models to generate text with citations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.
- Reshmi Ghosh, Rahul Seetharaman, Hitesh Wadhwa, Somyaa Aggarwal, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazadeh. 2024. [Quantifying reliance on external information over parametric knowledge during retrieval augmented generation \(RAG\) using mechanistic analysis](#). In *The 7th BlackboxNLP Workshop*.
- 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 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- John Hewitt. 2021. Initializing new word embeddings for pre-trained language models. <https://www.cs.columbia.edu/~johnhew/vocab-expansion.html>. Blog post.
- Hugging Face. 2025. Smollm3: smol, multilingual, long-context reasoner. <https://huggingface.co/blog/smollm3>. Hugging Face blog post.
- Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel Bell. 2026. [Abstentionbench: Reasoning LLMs fail on unanswerable questions](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Nikita Krayko, Ivan Sidorov, Fedor Laputin, Alexander Panchenko, Daria Galimzianova, and Vasily Konovalov. 2025. [RURAGE: robust universal RAG evaluator for fast and affordable QA performance testing](#). In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V*, Lecture Notes in Computer Science, pages 135–145. Springer.
- Pierre-Carl Langlais, Pavel Chizhov, Mattia Nee, Carlos Rosas Hinostroza, Matthieu Delsart, Irène Girard, Othman Hicheur, Anastasia Stasenko, and Ivan P. Yamshchikov. 2025a. [Even small reasoners should quote their sources: Introducing the pleias-rag model family](#). *CoRR*, abs/2504.18225.
- Pierre-Carl Langlais, Pavel Chizhov, Mattia Nee, Carlos Rosas Hinostroza, Matthieu Delsart, Irène Girard, Othman Hicheur, Anastasia Stasenko, and Ivan P. Yamshchikov. 2025b. [Even small reasoners should quote their sources: Introducing the pleias-rag model family](#). *CoRR*, abs/2504.18225.
- Bingbin Liu, Sébastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. 2023. [Tinygsm: achieving >80% on gsm8k with small language models](#). *CoRR*, abs/2312.09241.
- Ivan Maksimov, Vasily Konovalov, and Andrei Glinkii. 2024. [DeepPavlov at SemEval-2024 task 6](#).

- Detection of hallucinations and overgeneration mistakes with an ensemble of transformer-based models. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 274–278, Mexico City, Mexico. Association for Computational Linguistics.
- Maria Marina, Nikolay Ivanov, Sergey Pletenev, Mikhail Salnikov, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Alexander Panchenko, and Viktor Moskvoretskii. 2025. [Llm-independent adaptive RAG: let the question speak for itself](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 8697–8709. Association for Computational Linguistics.
- Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, H. Francis Song, Martin J. Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. [Teaching language models to support answers with verified quotes](#). *CoRR*, abs/2203.11147.
- Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. [Faitheval: Can your language model stay faithful to context, even if “the moon is made of marshmallows”](#). In *The Thirteenth International Conference on Learning Representations*.
- Kaixiang Mo, Yuxin Shi, Weiwei Weng, Zhiqiang Zhou, Shuman Liu, Haibo Zhang, and Anxiang Zeng. 2025. [Mid-training of large language models: A survey](#). *CoRR*, abs/2510.06826.
- Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. 2025. [Adaptive retrieval without self-knowledge? bringing uncertainty back home](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6355–6384, Vienna, Austria. Association for Computational Linguistics.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *CoRR*, abs/2508.10925.
- Sergey Pletenev, Maria Marina, Nikolay Ivanov, Daria Galimzianova, Nikita Krayko, Mikhail Salnikov, Vasily Konovalov, Alexander Panchenko, and Viktor Moskvoretskii. 2025a. [Will it still be true tomorrow? multilingual evergreen question classification to improve trustworthy QA](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8603–8620, Suzhou, China. Association for Computational Linguistics.
- Sergey Pletenev, Maria Marina, Daniil Moskovskiy, Vasily Konovalov, Pavel Braslavski, Alexander Panchenko, and Mikhail Salnikov. 2025b. [How much knowledge can you pack into a LoRA adapter without harming LLM?](#) In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4309–4322, Albuquerque, New Mexico. Association for Computational Linguistics.
- Gorjan Radevski, Kiril Gashteovski, Shahbaz Syed, Christopher Malon, Sebastien Nicolas, Chia-Chien Hung, Timo Sztyler, Verena Heußer, Wiem Ben Rim, Masafumi Enomoto, Kunihiro Takeoka, Masafumi Oyamada, Goran Glavas, and Carolin Lawrence. 2025. [On synthesizing data for context attribution in question answering](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 16929–16950. Association for Computational Linguistics.
- Elisei Rykov, Kseniia Petrushina, Maksim Savkin, Valerii Olisov, Artem Vazhentsev, Kseniia Titova, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025. [When models lie, we learn: Multilingual span-level hallucination detection with PsiloQA](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11663–11682, Suzhou, China. Association for Computational Linguistics.
- Timo Schick and Hinrich Schütze. 2021. [It’s not just size that matters: Small language models are also few-shot learners](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.
- Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. [Trusting your evidence: Hallucinate less with context-aware decoding](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.
- Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaie Khoei. 2025. [Agentic retrieval-augmented generation: A survey on agentic RAG](#). *CoRR*, abs/2501.09136.
- Kaiser Sun, Fan Bai, and Mark Dredze. 2026. [Task matters: Knowledge requirements shape LLM responses to context-memory conflict](#). In *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pages 4154–4176. Association for Computational Linguistics.
- Gemma Team. 2025a. [Gemma 3 technical report](#). *CoRR*, abs/2503.19786.
- Qwen Team. 2025b. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- 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.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. [Knowledge conflicts for LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2024. [Natural language reasoning, A survey](#). *ACM Comput. Surv.*, 56(12):304:1–304:39.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. [R-tuning: Instructing large language models to say ‘I don’t know’](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Jianguo Zhang, Tian Lan, Ming Zhu, Zuxin Liu, Thai Hoang, Shirley Kokane, Weiran Yao, Juntao Tan, Akshara Prabhakar, Haolin Chen, Zhiwei Liu, Yihao Feng, Tulika Manoj Awalganekar, Rithesh R. N., Zeyuan Chen, Ran Xu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, and 2 others. 2025. [xlam: A family of large action models to empower AI agent systems](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 11583–11597. Association for Computational Linguistics.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. [TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Packages and Hyperparameters

We used all pretrained base models and instruction-tuned baselines from their official Hugging Face repositories and report the exact model identifiers. For mid-training, we used Hugging Face/TRL supervised fine-tuning stack with accelerate for distributed training and Liger kernels for efficiency.

Hyperparameter	Qwen3-{0.6B, 1.7B}
Optimizer	AdamW
Peak learning rate	1×10^{-4}
LR schedule	cosine
Warmup ratio	0.03
Weight decay	0.01
Mixed precision	bf16
Max sequence length	6,144
Epochs	1
Per-device batch size	{32, 16}
Number of GPUs	8
Global batch size	{256, 128}
Distributed strategy	FSDP
Memory-saving kernel	Liger fused linear CE
Total training tokens	9×10^9
Wall-clock train time (hours)	{17, 28}
GPU type	NVIDIA H100 (80 GB)

Table 5: Mid-training hyperparameters for the best performing mid-trained models: Qwen3-0.6B-Base, Qwen3-1.7B-Base.

B Dataset Statistics

Category	QA Pairs	Tokens (B)
Overall Corpus		
Total	3,250,000	~8.0
Breakdown by Question Type		
Single-hop	2,780,000	7.76
Multi-hop (total)	471,000	0.27
– Single-context	263,000	–
– Multi-context	165,000	–
– Abstain	43,000	–
Token Distribution by Component		
<i>Single-hop Subset</i>		
Distractor contexts	–	~5.82
Gold contexts	–	~1.16
Reasoning chains	–	~0.78
<i>Multi-hop Subset</i>		
All contexts	–	~0.26
Reasoning/other	–	~0.01

Table 6: Summary Statistics of the QA Corpus

C QA Generation Prompts

Query:

```
<|query_start|> What is the name of the unit that the main character of the series featuring Karen Hayes is a member of?<|query_end|>
```

```
<|source_start|><|source_id|>1 Karen Hayes is a fictional character on the television program ... <|source_end|>
```

```
<|source_start|><|source_id|>2 Maxwell Greevey is a fictional character played by George Dzundza on NBC's ... <|source_end|>
```

```
<|source_start|><|source_id|>3 Jack Bauer is a fictional character and the lead protagonist of the Fox television series ...<|source_end|>
```

Response:

```
<|query_start|>
```

The question asks for the name of a specific unit. To find this, I need to: (1) identify the television series featuring the character Karen Hayes, (2) identify the main character of that series, and (3) determine the name of the unit that main character is a member of.

```
<|query_analysis_end|>
```

```
<|source_analysis_start|>
```

```
<|source_id|>1: states that Karen Hayes is a fictional character on the television program "24". This identifies the series.
```

```
<|source_id|>2: discusses Maxwell Greevey and "Law & Order"; irrelevant to Karen Hayes.
```

```
<|source_id|>3: states that Jack Bauer is the lead protagonist of "24" and that he often works as a member of the Counter Terrorist Unit (CTU).
```

```
<|source_analysis_end|>
```

```
<|reasoning_start|>
```

From <|source_id|>1, the series featuring Karen Hayes is "24". From <|source_id|>3, the main character of "24" is Jack Bauer, and Jack Bauer works as a member of the Counter Terrorist Unit (CTU). Therefore, the unit is the Counter Terrorist Unit (CTU).

```
<|reasoning_end|>
```

```
<|status_start|> ANSWERABLE <|status_end|>
```

```
<|answer_start|> Counter Terrorist Unit (CTU) <|answer_end|>
```

D Inference Code

```
import re
from transformers import
↳ AutoModelForCausalLM, AutoTokenizer

MODEL = "occ-ai/OCC-RAG-1.7B"
tok =
↳ AutoTokenizer.from_pretrained(MODEL)
model =
↳ AutoModelForCausalLM.from_pretrained(
    MODEL, torch_dtype="auto",
    ↳ device_map="auto")

question = ("Which country is the inventor
↳ of the telephone, Alexander Graham
↳ Bell, buried in?")

docs = [
    "Alexander Graham Bell was a
    ↳ Scottish-born inventor best known
    ↳ for patenting the first practical
    ↳ telephone.",
    "Bell died on August 2, 1922, at his
    ↳ estate Beinn Bhreagh, near
    ↳ Baddeck, Nova Scotia, and was
    ↳ buried there.",
    "Nova Scotia is a province on the east
    ↳ coast of Canada.",
]

text = tok.apply_chat_template([{"role":
↳ "user", "content":
↳ question}], documents=[{"text": d} for
↳ d in docs], tokenize=False,
↳ add_generation_prompt=True,
↳ enable_thinking=False)

inputs = tok([text],
↳ return_tensors="pt").to(model.device)
out = model.generate(**inputs,
↳ max_new_tokens=2048)
outs = out[0][inputs.input_ids.shape[1]:]
resp = tok.decode(outs,
↳ skip_special_tokens=False)
print(resp)

m = re.search(r"<|answer_start|>(.*)",
↳ resp, re.DOTALL)
print("Answer:", m.group(1).strip() if m
↳ else "")
```