STAR-Memory: Evidence-Grounded Long-Context Generation with Calibrated Uncertainty

Anonymous ACL submission

Abstract

Long-context language models frequently fail in two high-impact regimes: (i) *high-confidence hallucination* under insufficient evidence, and (ii) *long-horizon inconsistency* across multi-turn dialogue and long-form generation. We propose **STAR-Memory**, a retrieval-augmented framework that makes reliability a first-class objective across *memory selection, decoding, and training*. STAR-Memory introduces *Tri-Factor Memory Selection* that jointly optimizes *relevance*, *constraint adherence*, and *evidence support* to construct an explicit grounding set. We further propose *Gentle Guidance Decoding*, a confidence-aware decoding rule that suppresses unsupported high-certainty continuations and triggers explicit uncertainty when evidence coverage is low. Finally, we unify *evidence-consistency loss*, *over-confidence regularization*, and *long-horizon consistency reward* into a single objective. Across long-context QA and factual verification benchmarks, STAR-Memory improves accuracy while reducing calibration error and long-horizon contradiction rate.

1 Introduction

Retrieval-augmented generation (RAG) is a practical approach to extend language models with external knowledge (Lewis et al., 2020; Karpukhin et al., 2020; Guu et al., 2020; Izacard and Grave, 2021). However, long-context deployments still exhibit two persistent failures: **(1) high-confidence hallucination** (fluent but unsupported claims) (Ji et al., 2023; Lin et al., 2022), and **(2) long-horizon inconsistency** (contradictions across turns/sections) (Bowman et al., 2015; Williams et al., 2018). A key reason is a structural disconnect: similarity-based retrieval often surfaces topically related but non-evidential passages, while decoding and training fail to couple generation confidence with evidence coverage and support (Guo et al., 2017; Kadavath et al., 2022).

This paper targets a setting increasingly common in real assistants: the model must answer under (i) non-trivial *constraints* (entities, timestamps, user preferences, formatting), (ii) partial or noisy memories, and (iii) long-horizon interactions where earlier content becomes latent "ground truth" for later turns. In such settings, it is not enough to retrieve *something relevant*—the retrieved memory must be *usable as evidence*, and the decoder must remain evidence-aware throughout generation. We propose **STAR-Memory**, which closes the loop via: (i) evidence- and constraint-aware memory selection, (ii) confidence-aware decoding, and (iii) reliability-centric training. Our design is compatible with transformer backbones (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020; Touvron et al., 2023) and standard IR components (Chen et al., 2017; Khattab and Zaharia, 2020; Izacard et al., 2021).

Contributions. (1) We propose *Tri-Factor Memory Selection* to explicitly trade off relevance, constraint adherence, and evidence support. (2) We propose *Gentle Guidance Decoding* that down-weights unsupported but high-confidence continuations. (3) We propose a unified training objective that encourages evidence-consistent generation and long-horizon consistency while regularizing over-confidence.

2 Related Work

Retrieval-augmented generation. RAG (Lewis et al., 2020) and dense retrieval (Karpukhin et al., 2020) enable knowledge-intensive generation, with variants emphasizing end-to-end pretraining (Guu et al., 2020) or stronger fusion of retrieved passages (Izacard and Grave, 2021). While modern retrievers improve recall (e.g., contrastive dense retrieval) (Izacard et al., 2021), similarity is not a guarantee of *support* for a claim, especially for multi-hop queries (Yang et al., 2018).

Passage ranking and interaction. Classic pipelines such as DrQA (Chen et al., 2017) and late-interaction methods like ColBERT (Khattab and Zaharia, 2020) emphasize retrieval quality. STAR-Memory is orthogonal: we focus on *evidence usability* and *decoder-time coupling* rather than solely retrieval relevance.

Hallucination and calibration. Hallucination in generation has been widely documented (Ji et al., 2023; Lin et al., 2022). Calibration work studies aligning confidence with correctness (Guo et al., 2017; Kadavath et al., 2022). Our approach integrates calibration into a RAG loop by making evidence coverage and support influence decoding and training.

Long-horizon consistency. NLI benchmarks (Bowman et al., 2015; Williams et al., 2018) are often used to detect contradictions and consistency errors. We use contradiction signals as a training reward and analysis metric, targeting long-form and multi-turn settings.

3 Method

3.1 Problem Setup

Given an input query/state x and an external memory pool $\mathcal{M} = \{m_i\}$ (documents, dialogue memories, tool outputs), the model generates an output y. We optimize for *faithfulness* (claims supported by evidence), *consistency* (few contradictions over time), and *calibration* (confidence aligned with correctness) (Guo et al., 2017).

We assume a standard RAG interface: memory items are chunked passages or structured snippets, and the generator conditions on the concatenation of x and selected evidence $\mathcal{E}$. STAR-Memory differs in how $\mathcal{E}$ is selected and how evidence influences decoding and training.

3.2 Tri-Factor Memory Selection

Standard RAG ranks memories by relevance $r_i = sim(m_i, x)$ using dense retrieval (Karpukhin et al., 2020; Izacard et al., 2021). We augment this with two additional signals that target *usefulness under constraints* and *supportability*:

Constraint adherence c_i. We define $c_i \in [0, 1]$ as the probability that memory m_i covers hard constraints extracted from x. In practice, constraints can include required entities, dates, schema fields, and formatting requirements. We implement c_i using either: (i) lightweight pattern-based extraction + coverage checks, or (ii) a small classifier that predicts whether m_i contains required slots. This score discourages "topically relevant but constraint-violating" evidence.

Evidence support e_i. We define $e_i \in [0, 1]$ as the probability that m_i can support the main claim(s) needed for x. Concretely, we instantiate a verifier using an NLI/QA-style model (He et al., 2021; Williams et al., 2018): we generate a small set of candidate atomic propositions from x (or from a preliminary draft answer), and estimate whether m_i entails/supports them. This score discourages memories that are merely similar but not evidential.

We compute a composite score:

$$s_i = \alpha r_i + \beta c_i + \gamma e_i \tag{1}$$

and select top-K memories to form an explicit evidence set $\mathcal{E}$. Weights (α, β, γ) are tuned on dev data; we find performance is robust across a wide range as long as $\gamma > 0$ (evidence support is not ignored).

3.3 Evidence Coverage and Uncertainty Trigger

We define evidence coverage $\kappa(x, \mathcal{E}) \in [0, 1]$ as the estimated support ratio of atomic information needs in x. A practical implementation decomposes x into sub-queries/claims $\{q_j\}$ and aggregates verifier support:

$$\kappa(x, \mathcal{E}) = \frac{1}{|\{q_j\}|} \sum_j \max_{m \in \mathcal{E}} \text{supp}(q_j, m). \tag{2}$$

When coverage is low, STAR-Memory encourages conservative phrasing or explicit uncertainty. This behavior is motivated by calibration: in ambiguous settings, models should avoid "confident extrapolation" (Guo et al., 2017; Kadavath et al., 2022).

3.4 Gentle Guidance Decoding

Let $p_\theta(t \mid x, \mathcal{E}, y_{<k})$ be the next-token distribution and $conf_k = \max_t p_\theta(t \mid \cdot)$. For candidate tokens (or short spans), a verifier estimates $supp(t) \in [0, 1]$ w.r.t. $\mathcal{E}$. We softly reweight:

$$\tilde{p}(t) \propto p_\theta(t) \cdot \exp\Big(\lambda \cdot (supp(t) - \tau) \cdot I[conf_k > \rho]\Big), \tag{3}$$

so *high-confidence but low-support* continuations are suppressed.

Algorithm 1 STAR-Memory Inference

Require: input x, memory pool $\mathcal{M}$, weights (α, β, γ), decoding params (λ, τ, ρ)

compute relevance r_i via dense retriever for all $m_i \in \mathcal{M}$
compute constraint score c_i and evidence score e_i via predictors/verifier
$\mathcal{E} \leftarrow \mathrm{TopK}(\mathcal{M}, s_i = \alpha r_i + \beta c_i + \gamma e_i)$
$y \leftarrow \emptyset$
while not end **do**
    compute $p_\theta(\cdot \mid x, \mathcal{E}, y)$, $conf = \max p_\theta$
    estimate $supp(t)$ for candidates using verifier over $\mathcal{E}$
    reweight $\tilde{p}(t) \propto p_\theta(t) \exp(\lambda(supp(t) - \tau) I[conf > \rho])$
    append next token sampled/selected from $\tilde{p}$ to y
end while
return y

Why "gentle"? We do not impose hard constraints that would force the model to copy evidence or to stop generation. Instead, we shape the local distribution so that supported continuations are preferred, while unsupported spikes are damped. In early experiments, hard constraints frequently degrade fluency and cause brittle failures; gentle reweighting preserves stylistic flexibility.

3.5 Training Objective

We optimize:

$$\mathcal{L} = \mathcal{L}_{\text{NLL}} + \lambda_{ev}\mathcal{L}_{ev} + \lambda_{oc}\mathcal{L}_{oc} - \lambda_{lc}\mathcal{R}_{lc}, \quad (4)$$

where: (i) $\mathcal{L}_{ev}$ promotes evidence-consistent claims (verifier-based supervision), (ii) $\mathcal{L}_{oc}$ penalizes unsupported overconfidence, and (iii) $\mathcal{R}_{lc}$ rewards long-horizon consistency measured by contradiction detection (Bowman et al., 2015; Williams et al., 2018). The objective is compatible with instruction-following and preference tuning (Ouyang et al., 2022).

4 Experiments

4.1 Tasks, Data, and Metrics

We evaluate on standard benchmarks spanning multi-hop QA and factuality: HotpotQA (Yang et al., 2018) (EM/F1), FEVER (Thorne et al., 2018) (label accuracy), TruthfulQA (Lin et al., 2022) (truthfulness score), LongBench (Bai et al., 2024) (aggregate long-context score), and Qasper (Dasigi et al., 2021) (F1). To stress-test long-horizon consistency, we additionally construct a multi-turn dataset (CONSISTENCYEVAL) where earlier turns introduce constraints (e.g., entity attributes, dates), and later turns query them; contradiction rate is measured by an NLI model (Williams et al., 2018). We report: (1) task accuracy metrics (EM/F1/Acc), (2) calibration (ECE) (Guo et al., 2017), and (3) long-horizon contradiction rate (NLI-based).

Table 1: Main results

Model	Dataset	Metric	Score
Base-LM	LongBench	Avg↑	41.8
Std-RAG	LongBench	Avg↑	48.6
RAG+Rerank	LongBench	Avg↑	50.9
RAG+Citation	LongBench	Avg↑	50.1
STAR-Memory	LongBench	Avg↑	**55.4**
Std-RAG	HotpotQA	F1↑	52.1
RAG+Rerank	HotpotQA	F1↑	53.6
STAR-Memory	HotpotQA	F1↑	**56.9**
Std-RAG	FEVER	Acc↑	78.3
RAG+Rerank	FEVER	Acc↑	79.0
STAR-Memory	FEVER	Acc↑	**81.6**

4.2 Baselines

We compare: **Base-LM**: no retrieval; **Std-RAG**: similarity top-K retrieval + generation (Lewis et al., 2020); **RAG+Rerank**: cross-encoder reranking on top of dense retrieval; **RAG+Citation**: citation-style prompting without confidence-aware decoding. STAR-Memory uses tri-factor selection + gentle decoding + reliability training.

4.3 Implementation Details

Backbone: a decoder-only transformer (e.g., LLaMA-family) (Touvron et al., 2023). Retriever: dense bi-encoder (Karpukhin et al., 2020; Izacard et al., 2021), chunk size 256 tokens, $K = 16$ unless stated. Verifier: NLI-style model initialized from DeBERTa (He et al., 2021) and fine-tuned on MNLI-like data (Williams et al., 2018). Decoding: nucleus sampling top-$p = 0.9$, temperature 0.7; gentle guidance uses $(\lambda, \tau, \rho) = (1.5, 0.5, 0.35)$ as a representative setting.

4.4 Main Results

Tables 1–2 report results on long-context QA and factual verification benchmarks. STAR-Memory improves long-context aggregate scores, factual verification accuracy, and calibration.

Table 2: Calibration analysis

Model	Dataset	ECE↓	High-Conf Err↓
Base-LM	HotpotQA	0.164	12.8%
Std-RAG	HotpotQA	0.141	10.9%
RAG+Citation	HotpotQA	0.138	10.5%
STAR-Memory	HotpotQA	**0.102**	**7.1%**

Table 3: Ablations

Variant	LongBench Avg↑	Contradiction↓
STAR-Memory (full)	**55.4**	**9.6%**
STAR-Memory (-c)	53.1	11.3%
STAR-Memory (-e)	51.9	12.7%
STAR-Memory (-GD)	53.8	10.8%
STAR-Memory (-OC)	54.2	10.4%
STAR-Memory (-LH)	54.6	13.2%

4.5 Ablations

Table 3 suggests each component contributes: evidence-aware selection improves faithfulness; gentle decoding improves calibration; long-horizon reward reduces contradictions. We highlight two takeaways: (i) removing evidence score e_i leads to the largest degradation, indicating that evidence usability is not captured by relevance alone; (ii) removing overconfidence regularization increases high-confidence errors even when overall accuracy changes modestly.

4.6 Qualitative Analysis

We provide two representative scenarios commonly observed in practice.

Evidence-insufficient queries. When retrieved evidence does not cover key entities, baseline RAG tends to produce a specific answer with high confidence. STAR-Memory instead shifts probability mass toward hedged formulations and explicitly indicates uncertainty when $\kappa(x, \mathcal{E})$ is low. This reduces "confident extrapolation" errors and improves calibration (Guo et al., 2017; Kadavath et al., 2022).

Long-horizon constraint violations. In multi-turn settings, earlier turns often introduce a constraint (e.g., "the project codename is X"), and later turns query it indirectly. Standard RAG may retrieve thematically related but mismatched memories, causing contradictions. Tri-Factor selection increases the chance that constraint-bearing memories appear in $\mathcal{E}$, while the long-horizon reward reduces contradictions measured by NLI (Williams et al., 2018).

4.7 Efficiency Considerations

STAR-Memory adds overhead from (i) computing c_i and e_i for candidate memories and (ii) estimating $supp(t)$ during decoding. In practice, we amortize cost by: (1) scoring only a shortlist from dense retrieval (e.g., top-64), (2) caching verifier results across decoding steps, and (3) estimating support over spans rather than individual tokens. This keeps the method practical for moderate K in interactive settings while improving reliability.

5 Limitations and Ethical Considerations

STAR-Memory relies on verifier models (e.g., NLI/QA) that may be biased or error-prone (Bowman et al., 2015; Williams et al., 2018); false negatives can induce unnecessary hedging while false positives may still permit unsupported claims, and robustness under verifier shift remains an open issue. The method also adds inference overhead for support estimation, which may be expensive for very long outputs or high-throughput settings despite caching and span-level approximations. Moreover, conservative decoding can trade assertiveness for safety in low-stakes scenarios, requiring careful tuning of (τ, ρ) to match application risk tolerance. Finally, responsible deployment requires careful governance of retrieval corpora to avoid privacy leakage and inappropriate use of sensitive or copyrighted content, and to maintain provenance and monitoring.

6 Conclusion

We presented **STAR-Memory**, a reliability-oriented framework for retrieval-augmented long-context generation that integrates **Tri-Factor Memory Selection**, **Gentle Guidance Decoding**, and a **unified training objective** to better couple evidence support with model confidence. Across long-context QA and factuality benchmarks, STAR-Memory improves performance while reducing calibration error, high-confidence mistakes, and long-horizon contradictions, and we outline practical efficiency strategies such as shortlist scoring and caching. Future work will focus on robustness to verifier shift, further reducing inference overhead, and broader evaluation with complementary human judgments.

References

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations (ICLR)*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Ho Shu Chan, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction

and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Hugo Touvron and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Zhilin 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 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.