Epistemic State Modeling for Adaptive Retrieval-Augmented Generation

Nan Liang[1], Cheng Yang[2], Kun Yang[3], Jingqi Gao[1], Qingqiang Wu[1*], Meihong Wang[1]
[1] School of Informatics, Xiamen University, Xiamen, China
[2] School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China
[3] Institute of Artificial Intelligence, Xiamen University, Xiamen, China
liangnan@stu.xmu.edu.cn, youngcheng@my.swjtu.edu.cn, 36920241153270@stu.xmu.edu.cn
30920241154565@stu.xmu.edu.cn, wuqq@xmu.edu.cn, wangmh@xmu.edu.cn

***Abstract*—Retrieval-Augmented Generation (RAG) enhances LLMs with external knowledge, but existing methods use indiscriminate retrieval that ignores the model's internal knowledge state. We introduce Epistemic State Modeling (ESM), a framework that explicitly represents an LLM's metacognitive awareness as structured epistemic states over decomposed knowledge goals. Unlike prior adaptive RAG methods, ESM enables pre-generation planning by having the model confess what it knows, what it is uncertain about, and what specific gaps require retrieval. We implement ES-RAG via a comprehensive training methodology that jointly optimizes three metacognitive capabilities: (1) structured self-assessment via confession imitation, (2) confidence calibration via contrastive preference optimization, and (3) gap-driven intent generation for precise retrieval queries. Experiments on single-hop and multi-hop QA benchmarks show that ES-RAG achieves state-of-the-art performance while reducing retrieval overhead and providing transparent reasoning.**

***Index Terms*—Retrieval-Augmented Generation, Epistemic State, Adaptive Retrieval, Large Language Models**

I. Introduction

Large language models (LLMs) [1] have demonstrated remarkable capabilities in natural language understanding and generation, yet they remain fundamentally limited by their parametric knowledge—knowledge encoded during pre-training that becomes static post-deployment. This limitation manifests as factual hallucinations, outdated information, and poor performance on knowledge-intensive tasks [2]. Retrieval-Augmented Generation (RAG) [3] addresses these issues by dynamically incorporating external knowledge sources, enabling models to access up-to-date and domain-specific information beyond their training data.

However, standard RAG systems [3] operate under a critical assumption: all queries equally require external retrieval. This one-size-fits-all approach ignores a fundamental question—does the model actually need retrieval for this specific query? Recent work on adaptive RAG attempts to address this by introducing retrieval gating mechanisms, but these approaches face two key limitations. First, methods based on implicit signals (e.g., token probabilities [4], attention entropy [5]) lack semantic interpretability and cannot articulate what knowledge is missing. Second, methods employing explicit reflection (e.g., chain-of-thought prompting [6]) perform step-by-step reasoning during generation, but lack a structured representation of the model's epistemic state, missing the opportunity for pre-generation planning.

We argue that adaptive retrieval fundamentally requires the model to maintain an explicit representation of its *epistemic state*—a structured understanding of what it knows, what it is uncertain about, and what knowledge gaps exist. This metacognitive awareness should precede generation, enabling the model to plan retrieval actions based on identified knowledge deficits rather than merely reacting to generation failures.

To this end, we introduce **Epistemic State Modeling (ESM)**, a framework that treats adaptive RAG as a metacognitive planning problem. ESM decomposes each query into atomic knowledge goals and requires the model to confess its epistemic state for each goal before generation. This confession is structured along three dimensions: (1) confidence level (confident, uncertain, or unknown), (2) reasoning for the assessment, and (3) for uncertain or unknown goals, a precise specification of the knowledge gap and corresponding retrieval query.

We instantiate the ESM framework as **ES-RAG** (Epistemic State-aware RAG), an adaptive system that transforms retrieval from a reactive default into a proactive planning step. By explicitly modeling its own uncertainty prior to accessing external tools, ES-RAG selectively engages retrieval mechanisms only for specific, identified knowledge deficits, thereby improving efficiency while providing a fully interpretable trace of its metacognitive reasoning.

To cultivate reliable epistemic self-assessment, we develop a training methodology that targets the core challenges of metacognition. We employ *Structured Epistemic Induction (SEI)* to instill the capacity for structured self-reflection, while *Contrastive Metacognitive Calibration (CMC)* utilizes contrastive preference optimization to strictly align the model's confidence with its actual knowledge boundaries, effectively mitigating hallucination. Crucially, to transform passive uncertainty into active resolution, *Retrieval Utility Optimization (RUO)* empowers the model to articulate its specific information needs as precise retrieval intents.

Our contributions are threefold:

* indicates the corresponding author.

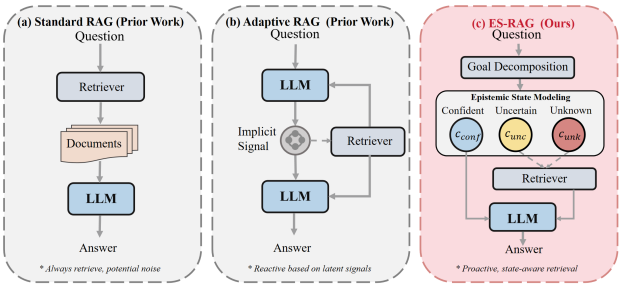

Fig. 1. Epistemic State–aware Retrieval Comparison. Standard RAG applies uniform retrieval; Adaptive RAG reacts to implicit uncertainty; ES-RAG proactively triggers selective retrieval via explicit epistemic state modeling.

- We formalize epistemic state modeling as a principled framework for adaptive RAG, shifting from reactive retrieval to proactive planning.
- We develop ES-RAG using a holistic training methodology that synergistically addresses the distinct challenges of metacognitive self-assessment, confidence calibration, and retrieval intent generation.
- We demonstrate through extensive experiments that ES-RAG achieves state-of-the-art performance on multiple open-domain QA benchmarks while substantially improving retrieval efficiency by selectively avoiding unnecessary retrieval and providing full interpretability of its reasoning process.

II. Related Work

A. Retrieval-Augmented Generation

RAG systems enhance LLMs by retrieving external knowledge before generation. Early methods like REALM [3] established this paradigm. Later work improved retrieval via iterative [7], multi-hop [8], or hybrid dense-sparse methods [9]. Yet, retrieval is applied to all queries, causing extra computation and potential noise when the model already knows the answer.

B. Adaptive Retrieval Mechanisms

Recognizing the inefficiency of universal retrieval, recent work explores adaptive mechanisms that selectively trigger retrieval.

Implicit signal-based methods monitor internal model states to detect knowledge deficiency. FLARE [10] uses token-level probability to trigger retrieval when generation confidence drops. DRAGIN [5] analyzes attention patterns to identify knowledge gaps. Self-RAG [4] trains special tokens to indicate retrieval necessity. While computationally efficient, these methods lack interpretability—they cannot explain why retrieval is needed or what specific knowledge is missing.

Explicit reflection-based methods use natural language reasoning to assess retrieval needs. Self-Ask [11] decomposes questions into sub-questions and retrieves for each. IRCoT [6] interleaves reasoning and retrieval in chain-of-thought generation. They are interpretable but perform reflection during generation and lack structured epistemic state representations.

C. Metacognition in Language Models

Our work connects to emerging research on metacognition in LLMs—the ability of models to reason about their own knowledge and reasoning processes. Studies have shown that LLMs can be prompted to express uncertainty and that their confidence correlates with accuracy [12]. However, recent work also reveals systematic miscalibration, with models exhibiting both overconfidence on incorrect answers and underconfidence on correct ones [13]. Our metacognitive calibration stage directly addresses this challenge through contrastive learning on counterfactual confessions. Unlike prior work that treats metacognition as a byproduct of prompting, we formalize it as a structured, trainable capability central to adaptive retrieval.

III. Methodology

A. Problem Formulation: Metacognitive Planning

We formalize adaptive Retrieval-Augmented Generation (RAG) not merely as a gating mechanism, but as a *metacognitive planning problem*. Given a user query q and an external knowledge corpus $\mathcal{D}$, the objective is to synthesize an answer a that maximizes factual accuracy while minimizing retrieval overhead. This requires the model to perform a pre-generation introspection to determine the necessity and scope of external information.

We define the *epistemic state space* $\mathcal{S}$ over a decomposed set of atomic knowledge goals $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$. For each goal g_i, the model must explicitly predict a state tuple $s_i = (c_i, r_i, k_i)$, where:

- $c_i \in \{\text{CONFIDENT}, \text{UNCERTAIN}, \text{UNKNOWN}\}$ denotes the calibrated confidence level regarding the model's parametric knowledge.
- r_i is a natural language rationale that explains the confidence assessment, reflecting the model's internal reasoning process.
- $k_i = (d_{\text{gap}}, q_{\text{ret}})$ represents the knowledge gap specification (d_{gap}) and the corresponding search query (q_{ret}), generated only when $c_i \neq \text{CONFIDENT}$.

The adaptive retrieval policy π_{ret} is thus deterministic and sparse: retrieval is executed for a goal g_i if and only if its epistemic state dictates external verification. This formulation shifts the paradigm from implicit uncertainty estimation (e.g., entropy thresholds) to explicit, interpretable metacognitive planning.

B. Inference Architecture

As illustrated in Figure 2, we instantiate this framework as ES-RAG (Epistemic State-aware RAG). The inference pipeline is strictly stratified to prevent "lazy retrieval"—the tendency of models to over-rely on context.

Phase 1: Pre-Retrieval Epistemic Confession. The model processes the query q under a strict constraint: access to external tools is blocked. It must generate a structured confession $\mathcal{C} = \{(g_i, s_i)\}_{i=1}^{n}$ relying solely on its pre-trained parametric

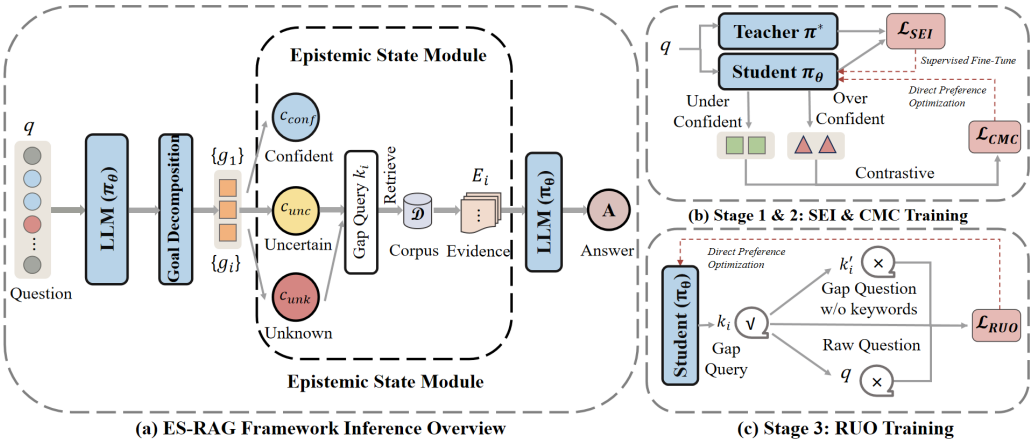

Fig. 2. The ES-RAG Inference Architecture. The process is stratified into four phases: (1) Pre-Retrieval Epistemic Confession, where the model introspects without external access; (2) Gap-Driven Retrieval, targeting only specific deficits; (3) Evidence-Grounded Synthesis; and (4) Interpretable Output generation.

memory. This forces the model to expose its true internal knowledge boundaries.

Phase 2: Gap-Driven Retrieval. For the subset of goals $\mathcal{G}_{\text{miss}} = \{g_i \mid c_i \in \{\text{UNCERTAIN}, \text{UNKNOWN}\}\}$, the system executes the generated queries $\{q_{\text{ret}}^{(i)}\}$ against the corpus $\mathcal{D}$, retrieving evidence sets $\mathcal{E} = \{E_i\}$. This enables surgical information acquisition, contrasting with the coarse-grained, query-level retrieval of standard RAG.

Phase 3: Evidence-Grounded Synthesis. The final answer a is generated by conditioning on both the original query and the sparse evidence $\mathcal{E}$. The prompt enforces a "bridge-and-fill" strategy: confident parametric knowledge is retained, while uncertain segments are grounded in $\mathcal{E}$.

C. Holistic Metacognitive Alignment

To endow the Large Language Model (LLM) with reliable epistemic capabilities, we propose a multi-stage, multi-task alignment framework targeting three complementary competencies—*Structured Epistemic Induction*, *Contrastive Metacognitive Calibration*, and *Retrieval Utility Optimization*—which support structured reasoning, confidence calibration, and effective retrieval query generation.

1) Structured Epistemic Induction (SEI): The foundational objective is to instill the capability to perform structured introspection. We utilize a high-fidelity teacher model (e.g., GPT-4o [1]) to synthesize a dataset $\mathcal{D}_{\text{syn}} = \{(q, \mathcal{C}^*)\}$, where $\mathcal{C}^*$ represents the ground-truth epistemic confession containing goals, confidence labels, and reasoning traces. We optimize the student model π_θ using standard Supervised Fine-Tuning (SFT) to minimize the negative log-likelihood of the structured confession tokens y:

$$\mathcal{L}_{\text{SEI}}(\theta) = -\mathbb{E}_{(q,y)\sim\mathcal{D}_{\text{syn}}}\left[\frac{1}{|y|}\sum_{t=1}^{|y|}\log \pi_\theta(y_t \mid q, y_{<t})\right] \quad (1)$$

This objective establishes the model's ability to decompose queries and adhere to the rigorous output schema required for downstream planning.

2) Contrastive Metacognitive Calibration (CMC): A critical pathology in LLMs is *epistemic misalignment*—the disconnect between subjective confidence and objective accuracy (e.g., hallucinated confidence [13]). To address this, we employ Direct Preference Optimization (DPO, [14]) to enforce calibration. For a given query q, we construct preference pairs (y_w, y_l) based on the correctness of the model's parametric knowledge:

- **Winning Response (y_w):** The *Accurate Confession*, where the confidence label aligns with ground truth (e.g., claiming CONFIDENT when the answer is correct, or UNCERTAIN when incorrect).
- **Losing Response (y_l):** The *Miscalibrated Confession*, representing either overconfidence (claiming CONFIDENT on errors) or underconfidence (claiming UNCERTAIN on known facts).

The calibration loss ensures the model assigns higher implicit reward to calibrated states:

$$\mathcal{L}_{\text{CMC}}(\theta) = -\mathbb{E}_{(q,y_w,y_l)\sim\mathcal{D}_{\text{cal}}}\log\sigma\Big(\beta_{\text{cal}}\big[\Delta_\theta^{\text{cal}} - \Delta_{\text{ref}}^{\text{cal}}\big]\Big) \quad (2)$$

where $\mathcal{D}_{\text{cal}}$ denotes the calibration preference dataset, β_{cal} controls the deviation penalty, π_{ref} is the reference policy derived from the SEI phase, $\Delta_\theta^{\text{cal}} = \log \pi_\theta(y_w \mid q) - \log \pi_\theta(y_l \mid q)$, and $\Delta_{\text{ref}}^{\text{cal}}$ is defined analogously using the reference policy π_{ref}.

This objective effectively penalizes blind confidence and encourages the model to acknowledge its limitations.

3) Retrieval Utility Optimization (RUO): Even with accurate self-diagnosis, the system fails if it cannot articulate *how* to resolve a knowledge gap. The final objective optimizes the quality of the generated retrieval queries (q_{ret}). We construct a preference dataset $\mathcal{D}_{\text{util}}$ containing pairs of query formulations for the same knowledge gap:

- **Winning Query (q^+):** A specific, keyword-optimized query that retrieves high-relevance evidence (verified by retriever recall).

- **Losing Query (q^-):** A query that is either overly broad (generic semantic drift) or overly narrow (keyword hallucination).

We apply DPO to optimize the retrieval intent generation:

$$\mathcal{L}_{\text{RUO}}(\theta) = -\mathbb{E}_{(c_{\text{ctx}},q^+,q^-)\sim\mathcal{D}_{\text{util}}} \log \sigma\Big(\beta_{\text{util}}\big[\Delta_\theta^{\text{util}} - \Delta_{\text{ref}}^{\text{util}}\big]\Big) \quad (3)$$

where c_{ctx} denotes the context of the identified knowledge gap, β_{util} is a scaling factor controlling the strength of the preference signal between winning and losing retrieval queries, $\Delta_\theta^{\text{util}} = \log \pi_\theta(q^+ \mid c_{\text{ctx}}) - \log \pi_\theta(q^- \mid c_{\text{ctx}})$, and $\Delta_{\text{ref}}^{\text{util}}$ is defined analogously using the reference policy π_{ref}.

This completes the alignment loop, ensuring that the model's metacognitive awareness translates into effective retrieval actions.

IV. Experiments

A. Experimental Setup

Datasets. We evaluate ES-RAG on four open-domain question answering (QA) benchmarks that vary in reasoning complexity:

- **Single-Hop QA**: We use **Natural Questions** (NQ, [15]) and **TriviaQA** [16]. These datasets primarily test the model's precision in retrieving and utilizing factual knowledge.
- **Multi-Hop QA**: We use **HotpotQA** [17] and **2WikiMultiHopQA** (2Wiki, [18]). These require aggregating information from multiple documents, rigorously testing the model's planning and gap-driven retrieval capabilities.

We sampled questions from each benchmark and generated 34,237 structured synthetic training instances using GPT-4o, following the ES-RAG knowledge gap formulation. For evaluation, 2,000 instances were randomly drawn from each dataset, ensuring no overlap with the synthetic training data.

Implementation Details. We use the Dec 20, 2018 English Wikipedia with BM25 [19] as retriever and LLaMA-2-chat-7B [20] as backbone. Training follows our holistic alignment framework with LoRA (Rank=8, Alpha=16) in bf16, 2 epochs, batch size 2, LR $5e^{-5}$, and overconfidence sampling $p_{\text{cal}} = 0.5$ for Contrastive Metacognitive Calibration.

Baselines. We compare ES-RAG against three categories of baselines, including the most recent state-of-the-art (SOTA) adaptive methods:

- **Standard LLMs**: *CoT* [8] and *Vanilla LLM* without retrieval.
- **Standard RAG**: *ReAct* [21] and *Self-Ask* [11].
- **Adaptive RAG**: *FLARE* [10] (confidence-based active retrieval), *Self-RAG* [4] (reflection tokens), *DRAGIN* [5] (adaptive retrieval on low-confidence tokens), *Adaptive-RAG* [22] (complexity-aware adaptive retrieval), and *SEAKR* [23] (self-aware knowledge retrieval).

Evaluation Metrics. We report Exact Match (EM) and F1 score for answer quality.

B. Main Results

Table I summarizes the performance comparison. ES-RAG consistently outperforms all baselines across both single-hop and multi-hop datasets.

Superiority in Multi-Hop Reasoning: On complex datasets like HotpotQA and 2WikiMultiHop, ES-RAG achieves F1 scores of 45.6% and 39.0%, respectively, surpassing the strong baseline Adaptive-RAG (45.3% on HotpotQA, 38.1% on 2Wiki). This validates the effectiveness of our *Structured Epistemic Induction (SEI)*, which enables the model to explicitly decompose complex queries into manageable knowledge goals, rather than treating retrieval as a flat driven process without explicit gap modeling (as in Standard RAG) or relying on implicit signals (as in FLARE).

Calibration Improves Precision: On TriviaQA, ES-RAG achieves a remarkable 71.4% F1, significantly higher than Self-RAG (46.4%). We attribute this to the *Contrastive Metacognitive Calibration (CMC)*. TriviaQA contains many obscure facts; baselines often hallucinate or retrieve irrelevant context. ES-RAG's calibrated confidence ensures it only retrieves when truly necessary, reducing noise injection.

C. Ablation Study: The Role of Alignment Objectives

To verify the contribution of each component in our holistic alignment framework, we conducted an ablation study (Table II). We remove each optimization objective while keeping others constant.

- **w/o SEI**: Removing the foundational structure learning (formerly Stage 1) leads to the most drastic performance drop (TriviaQA F1 -11.3%). This confirms that the ability to generate structured "epistemic confessions" is the prerequisite for adaptive planning.
- **w/o CMC**: Removing the calibration objective (formerly Stage 2) results in a notable drop, particularly on HotpotQA (-6.0% F1). Without calibration, the model suffers from overconfidence propagation, failing to verify intermediate steps in multi-hop reasoning.
- **w/o RUO**: Removing the intent optimization (formerly Stage 3) degrades performance by preventing the generation of search-friendly queries, proving that "knowing what you don't know" is insufficient without "knowing how to ask".

D. Retrieval Efficiency and Cost

A key promise of adaptive RAG is resource efficiency. We analyze the average number of Retrieval Calls (RC) and Language Model Calls (LMC) per query on the HotpotQA set (Figure 3).

ES-RAG achieves a strategic balance. While simple baselines (Naive RAG) use fixed resources, advanced adaptive methods like DRAGIN and SEAKR incur high computational costs (up to 9.9 LM calls). ES-RAG maintains high accuracy with only **1.52 retrieval calls** and **2.42 LM calls** on average. Importantly, this average masks a bimodal behavior. On the HotpotQA set, we observe that 28.8% of queries are classified

TABLE I
COMPARATIVE PERFORMANCE OF ES-RAG AGAINST STATE-OF-THE-ART BASELINES ON FOUR OPEN-DOMAIN QA BENCHMARKS.

Method	NQ		TriviaQA		2Wiki		HotpotQA	
	F1	EM	F1	EM	F1	EM	F1	EM
No Retrieval								
Vanilla LLM	21.2	16.1	44.8	38.5	18.4	13.2	24.1	15.9
CoT	23.7	18.4	46.6	40.8	21.2	15.3	26.6	16.8
Standard & Adaptive Retrieval								
ReAct	27.4	20.1	53.2	46.0	28.0	21.0	41.7	24.9
FLARE	35.9	25.3	60.3	51.5	21.3	14.3	22.1	14.9
Self-Ask	31.6	24.3	57.8	49.6	37.5	28.6	43.1	28.2
Self-RAG	40.2	32.3	46.4	37.1	29.5	21.8	33.4	22.9
DRAGIN	33.2	23.2	62.3	54.0	30.0	22.4	34.2	23.7
Adaptive-RAG	45.8	36.9	59.3	53.0	38.1	32.4	45.3	37.0
SEAKR	35.5	25.6	63.1	54.4	36.0	30.2	39.7	27.9
ES-RAG (Ours)	**46.4**	**40.8**	**71.4**	**66.0**	**39.0**	**34.1**	**45.6**	**39.8**

TABLE II
ABLATION ANALYSIS OF OPTIMIZATION OBJECTIVES.

Configuration	TriviaQA		HotpotQA	
	F1	EM	F1	EM
ES-RAG (Full)	**71.4**	**66.0**	**45.6**	**39.8**
w/o SEI (Structure)	60.1	54.9	34.1	30.1
w/o CMC (Calibration)	69.2	63.8	39.6	35.6
w/o RUO (Utility)	70.7	65.3	43.8	39.2

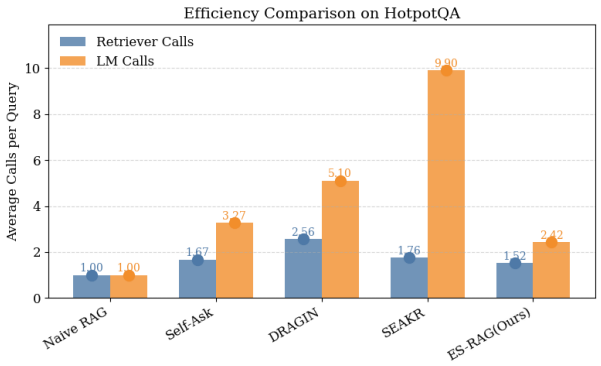

Fig. 3. Efficiency comparison on HotpotQA. RC: Retriever Calls/Query; LMC: LM Calls/Query.

as c_i = CONFIDENT by ES-RAG, for which no retrieval is triggered (i.e., zero retrieval calls).

E. Impact of Calibration Hyperparameters

We study CMC's sensitivity to the overconfidence rate p_{cal}, which controls the probability of sampling overconfident negatives, on TriviaQA. Losing responses y_l are sampled as overconfident (CONFIDENT on errors) or underconfident (UNCERTAIN on correct answers). Figure 4 shows $p_{\text{cal}} = 0.5$

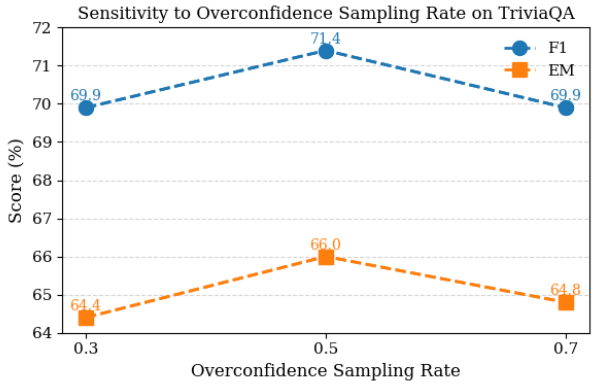

Fig. 4. Sensitivity to overconfidence sampling rate (p_{cal}) on TriviaQA.

is optimal; lower values under-penalize hallucinations, higher values make the model overly conservative, increasing retrieval and latency.

F. Case Study: Interpretable Planning

Table III illustrates the interpretable reasoning of ES-RAG compared to Naive RAG. Faced with a complex multi-hop question about an album's guest artist, Naive RAG is distracted by irrelevant documents (Britney Spears). In contrast, ES-RAG exhibits a structured and transparent planning process:

- **Uncertainty Identification.** The model recognizes that while it is familiar with the album, it lacks confidence regarding the specific guest artists, and explicitly marks the corresponding goal as UNCERTAIN.
- **Plan Formulation.** Based on the identified knowledge gap, ES-RAG generates a targeted retrieval query focusing on *guest appearances*, enabling evidence-driven reasoning.

TABLE III
CASE STUDY: HANDLING KNOWLEDGE GAPS.

Query: *Fast Cars, Danger, Fire and Knives includes guest appearances from which hip hop record executive?*
Naive RAG: Retrieves documents about Britney Spears (irrelevant). **Output**: Britney Spears ✗
ES-RAG (Ours): **Goal 1**: Identify the album's basic information. → *Status*: CONFIDENT. **Goal 2**: Identify guest artists. → *Status*: UNCERTAIN. *Gap Query*: "Fast Cars... guest appearances" *Evidence*:...features Camu Tao and **El-P**... **Goal 3**: Confirm if El-P is an executive. → *Status*: UNKNOWN. *Gap Query*: "El-P Definitive Jux executive" *Evidence*: El-P... is CEO of Definitive Jux... **Output**: Jaime Meline (El-P) ✓

- **Recursive Knowledge Refinement.** After retrieving evidence related to *El-P*, the model initiates a second verification step to determine whether *El-P* qualifies as a *hip hop record executive*. This recursive validation demonstrates dynamic, self-correcting planning behavior rather than one-shot retrieval.

V. CONCLUSION

We introduced ES-RAG, a proactive metacognitive planning framework that models epistemic states to bridge parametric memory and non-parametric retrieval, achieving SOTA on four benchmarks, with future work targeting multimodal retrieval, efficiency, and AI systems that know their limits.

ACKNOWLEDGMENT

This work is supported by the Solfeggio ear training intelligent robot and cloud platform research and development project for music education (No.2024CXY0102), the 3D visualization digital twin integrated control system (No.2023CXY0111), the public technology service platform project of Xiamen City (No.3502Z20231043) and Fujian Provincial Science and Technology Major Project (No.2024HZ022003).

REFERENCES

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[3] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *International conference on machine learning*. PMLR, 2020, pp. 3929–3938.

[4] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-rag: Learning to retrieve, generate, and critique through self-reflection," 2024.

[5] W. Su, Y. Tang, Q. Ai, Z. Wu, and Y. Liu, "DRAGIN: Dynamic retrieval augmented generation based on the real-time information needs of large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Aug. 2024, pp. 12 991–13 013.

[6] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions," in *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, 2023, pp. 10 014–10 037.

[7] Z. Shao, Y. Gong, Y. Shen, M. Huang, N. Duan, and W. Chen, "Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds., Dec. 2023.

[8] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[9] X. Ma, Y. Gong, P. He, H. Zhao, and N. Duan, "Query rewriting in retrieval-augmented large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 5303–5315.

[10] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, and G. Neubig, "Active retrieval augmented generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 7969–7992.

[11] O. Press, M. Zhang, S. Min, L. Schmidt, N. A. Smith, and M. Lewis, "Measuring and narrowing the compositionality gap in language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 5687–5711.

[12] M. Joglekar, J. Chen, G. Wu, J. Yosinski, J. Wang, B. Barak, and A. Glaese, "Training llms for honesty via confessions," *arXiv preprint arXiv:2512.08093*, 2025.

[13] Y. Li, K. Zhou, Q. Qiao, B. Nguyen, Q. Wang, and Q. Li, "Investigating context faithfulness in large language models: The roles of memory strength and evidence style," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 4789–4807.

[14] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in neural information processing systems*, vol. 36, pp. 53 728–53 741, 2023.

[15] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee *et al.*, "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019.

[16] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1601–1611.

[17] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2369–2380.

[18] X. Ho, A.-K. Duong Nguyen, S. Sugawara, and A. Aizawa, "Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Dec. 2020.

[19] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[20] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[21] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *The eleventh international conference on learning representations*, 2022.

[22] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. Park, "Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Jun. 2024, pp. 7036–7050.

[23] Z. Yao, W. Qi, L. Pan, S. Cao, L. Hu, L. Weichuan, L. Hou, and J. Li, "SeaKR: Self-aware knowledge retrieval for adaptive retrieval augmented generation," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jul. 2025.