Transactions on Computational and Scientific Methods | Vo. 5, No. 11, 2025
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

LLM Retrieval-Augmented Generation with Compositional Prompts and Confidence Calibration

Song Han
Northeastern University, Boston, USA
song.han.dev@gmail.com

Abstract: This study investigates the reliability of Retrieval-Augmented Generation (RAG) in complex and dynamic knowledge settings, and proposes a compositional retrieval-prompting framework with gated knowledge injection and layered confidence calibration. The method first performs semantic parsing and prompt decomposition to transform complex queries into structured expressions composed of sub-intents and logical operators, providing a clear planning path for the retrieval stage. In the knowledge injection process, gating and filtering mechanisms are introduced to effectively suppress noisy fragments and enhance the relevance and controllability of evidence, allowing retrieval results to align more accurately with the generation model. During generation, the model applies multi-granularity evidence fusion strategies to optimize answers and measures uncertainty on both the retrieval and generation sides within a layered confidence calibration framework, ensuring traceability and consistency of outputs. Systematic experiments on hyperparameter sensitivity, environmental constraints, and data transfer show that the framework demonstrates strong robustness and stability across different scenarios, significantly improving answer accuracy, factual consistency, and attribution ability under complex task conditions, thereby providing an effective solution for knowledge-intensive generation tasks.

Keywords: Retrieval-enhanced generation; compositional hints; knowledge injection; confidence calibration.

1. Introduction

Large generative models show strong abilities in open-domain question answering, enterprise knowledge services, and decision support. Their parametric memory, however, lags behind the dynamic evolution of external knowledge. This often leads to hallucinations and uncertain responses on cross-domain transfer, time-sensitive facts, and long-tail entities. Retrieval augmented generation introduces external corpora to supply traceable evidence. It offers a feasible path to improve factual consistency and explainability. Yet the traditional pattern of single-round retrieval with concatenated prompts becomes fragile under massive scale and heterogeneous corpora. Common issues include query granularity mismatch, insufficient evidence coverage, and redundant injection. These problems limit knowledge alignment and reliability[1]. In this work, we address these limitations with a compositional retrieval-prompting framework that decomposes complex queries into operator-based sub-intents, injects evidence under dynamic gating, and jointly calibrates retrieval- and generation-side signals.

The core difficulty of knowledge injection is not only finding the right information but also injecting it correctly. Knowledge from different sources, times, and structural forms carries risks of conflict, noise, and

staleness. Simple fragment concatenation before generation induces semantic shortcuts, attention dilution, and unclear evidence attribution. Knowledge injection, therefore, requires coordinated design along three dimensions. These are semantic alignment, structural consistency, and controllability with strong constraints. The goal is to reduce the mismatch between query intent and evidence granularity. It is also to maximize the density of effective information under input budget limits. A traceable attribution mechanism must ensure that conclusions are causally verifiable from the evidence. This shift toward injecting to guide generation implies that retrieval, prompting, and generation should not be treated in isolation. They should be viewed as a unified process of knowledge flow and constraint propagation[2,3].

Compositional retrieval prompting provides the infrastructure for complex intent decomposition and evidence organization. Compared with a single query, a compositional prompt decomposes a compound task into composable sub-intents and operators[4]. It supports multi-hop, multi-view, and multi-constraint retrieval planning. This improves evidence coverage and relevance. The compositional structure also fits hierarchical alignment and stepwise convergence control. It reduces prompt fragility and template dependence while preserving interpretability. In practice, heuristic manual prompt assembly is hard to reuse with stability. Small linguistic changes can shift the retrieval distribution and degrade evidence quality. An algorithmic prompt-driven mechanism is therefore needed. Semantic parsing, operator composition, and retrieval feedback should form a closed loop. This enables fine-grained scheduling of knowledge injection and robust deployment[5].

Confidence calibration is a trust backbone across retrieval, reranking, injection, and generation. Real corpora suffer from sparse labels, uneven sources, and temporal drift. Retrieval scores, reranking scores, and generation probabilities often live on different scales and do not match. The result is systematic bias, including high-confidence errors and low-confidence correct answers. Poor calibration also hides evidence of conflicts and sources of uncertainty[6]. It weakens triggers for safety actions such as refusal, follow-up questioning, and reporting diverse candidates. A layered confidence representation and joint calibration are needed at the passage level, the claim level, and the answer level. The model can then quantify knowledge reliability. It can choose responses under risk control, weigh evidence, and switch strategies. This creates measurable trust boundaries for compliance and high-stakes scenarios[7].

Research on knowledge injection and confidence calibration driven by compositional retrieval prompting has clear theoretical and practical value. On the theoretical side, it can unify retrieval, prompting, and generation within a framework of composable semantics and probabilistic calibration. It can describe how knowledge constraints propagate through prompt structures and shape the generative distribution. It advances the field from an empirical view where relevance is treated as sufficient to a mechanism-based view that is evidence-based and confidence calibrated. On the practical side, it can improve factual consistency, attribution transparency, and risk controllability with manageable cost. It supports deployable knowledge governance in domains with rapid updates, strict compliance, and low tolerance for risk. It also lays a scalable foundation for verifiable generation, traceable decision making, and cross-domain transfer[8].

2. Related work

Existing research on retrieval augmented generation largely follows a retrieve-then-generate paradigm. In particular, it focuses on index construction, chunking strategies, fusion of dense and sparse retrieval, reranking, and evidence fusion. Mainstream systems use semantic encoders to strengthen cross-domain matching. They combine windowed chunking with hierarchical recall to expand evidence coverage. In the generation stage, they concatenate context and enforce citation constraints to improve traceability. However, as the knowledge base grows in scale and heterogeneity, a serise of structural challenges emerge including granularity mismatch, knowledge conflicts, attention dilution in long contexts, and tight input budgets. A

metric mismatch between retrieval scores and generation confidence also appears. It biases evidence weighting and answer selection. Many studies, therefore, explore joint optimization and closed-loop feedback across retrieval, reranking, and generation. Even so, evidence organization and constraint propagation remain limited under complex intents, multi-hop reasoning, and time-varying knowledge[9]. The RAG formulation introduces retrieve-then-generate with non-parametric memory [10]. DPR improves passage recall with dense dual-encoders [11], and FiD aggregates evidence by independently encoding passages and fusing them in the decoder [12]. Building on these, RAG-end2end targets domain adaptation [13], while REAR explicitly assesses document relevance for safer utilization of retrieved knowledge [14]. For evaluation, eRAG provides document-level retrieval quality aligned with downstream RAG performance rather than serving as a generative baseline [15].

Compositional prompting addresses prompt engineering and task decomposition. It splits complex queries into composable subgoals and operators. Typical operators include entity constraints, temporal constraints, logical connectives, and evidence aggregation. Planned retrieval and progressive constraints then narrow the evidence space. This improves relevance and coverage. The approach expresses intent with an interpretable structure. It supports multi-view alignment, constraint stacking, and path pruning. It fits multi-hop retrieval, temporal constraints, and domain restrictions. Compared with fixed templates or single round prompts, compositional structure offers potential gains in robustness and transferability. Nevertheless, current practice still relies on heuristic assembly and offline rules. Small wording changes and retrieval noise often amplify distribution shift and quality variance. The feedback from retrieval rarely forms a learnable closed loop with the prompt structure. Policies for when to refine, where to stop, and how to merge do not generalize well. A key gap is an algorithmic compositional prompt that drives retrieval planning and is jointly constrained with downstream generation objectives[16]. RCR retrieves step-by-step with a tri-encoder and reinforcement learning, composing informative contexts under an MDP view [17]. GraphRAG constructs a corpus-level graph (entities/links) and plans multi-hop expansion over the graph structure [18]. In parallel, Self-RAG lets an LM decide when to retrieve and how to critique via reflection tokens [19].

On knowledge injection, research now covers what to take, how to place it, and under what constraints. On the retrieval side, it studies index refresh, domain adaptation, choices of chunking and aggregation scale, and cross-corpus deduplication with conflict detection. On the bridging side, it uses contrastive learning or multi-objective reranking for evidence selection, claim-level aggregation, and redundancy suppression. On the generation side, it controls the injection position, the encoding of control signals, controllable decoding, and citation consistency constraints. Some work explores joint training of retrieval and generation, learnable evidence gating, and consistency regularization between answers and evidence. These aim to reduce hallucinations and improve attribution[20]. RAGate predicts whether a turn needs external knowledge and gates low-utility passages [21]; CAG (Context Awareness Gate) dynamically adjusts prompts based on query context to suppress noisy injection [22]. Beyond retrieve-then-generate, GenGround alternates generate-then-ground steps to correct earlier errors with retrieved evidence in multi-hop settings [23]. Despite these advances, injection granularity is often decoupled from the prompt structure. Intent, evidence, and answer lack consistent structural alignment. Handling of evidence conflicts and temporal drift remains mostly heuristic. Stable policy learning and verifiable attribution paths are still limited.

Research on confidence and uncertainty provides trust guarantees for retrieval augmented generation. Existing methods measure generative uncertainty using posterior probabilities, energy scores, entropy, and distributional distances. A recent survey systematizes LLM confidence estimation/calibration and gaps between verbalized confidence and correctness; SelfCheckGPT detects hallucinations via self-consistency probing without model internals [24]. At fact level, granular calibration aligns confidence with per-claim correctness and supports self-correction. A complementary RAG evaluation survey reviews retrieval-side

and generation-side metrics (relevance, accuracy, faithfulness) and their correlations [25]. They reduce calibration error using temperature scaling, monotone piecewise mapping, and hierarchical calibration. Other studies estimate confidence at the passage, claim, and source levels on the retrieval and reranking sides. These estimates support evidence weighting and trigger refusal or follow-up questioning. New work stresses multi-source uncertainty alignment across stages. It seeks to place retrieval similarities, reranking scores, and generation probabilities on a common scale. This helps detect and correct high-confidence errors and low-confidence correct answers. Even so, calibration is not yet integrated throughout compositional prompting and knowledge injection. Decomposition and merging in the prompt can shift evidence distributions and alter decoding boundaries. This harms the stability of confidence estimates. In real-time and cross-domain settings, explainable risk control and strategy switching are still not systematic. A unified framework that ties compositional retrieval prompting, knowledge injection, and layered confidence calibration is therefore needed. Such a framework is central to moving retrieval augmented generation from relevance to trust[26].

3. Method

This study introduces an algorithmic approach for knowledge injection and confidence calibration driven by compositional retrieval prompting. The core idea is to decompose complex query intents into composable retrieval subunits, use structured prompts to achieve layered injection of external knowledge, and establish a unified confidence representation and calibration mechanism across retrieval, prompting, and generation. The framework has three main stages. First, a semantic parsing and prompt composition module converts the input query into a structured expression composed of semantic subgoals and logical operators. Second, a knowledge injection mechanism selects and maps evidence through multi-granularity alignment and dynamic gating. Finally, a multi-level confidence calibration model provides a unified measurement of retrieval scores, generation probabilities, and structural consistency. This improves system reliability and controllability in dynamic knowledge environments. The model architecture is shown in Figure 1.

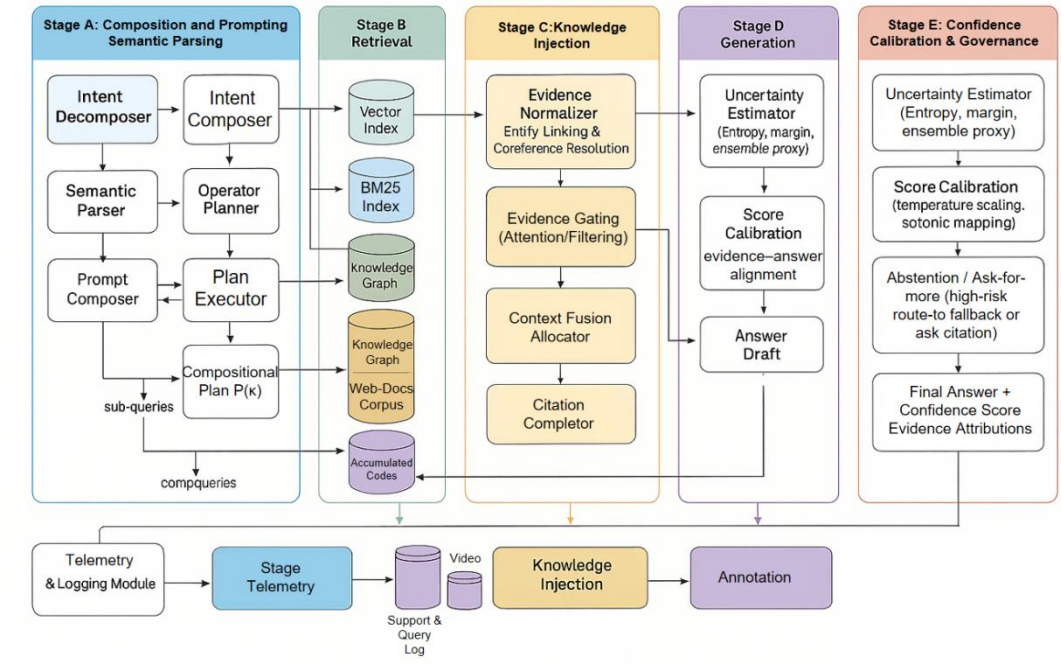

Figure 1. RAG knowledge injection and confidence calibration framework based on combined retrieval

In the compositional hint generation phase, the input query sequence $x = \{x_1, x_2, ..., x_n\}$ is first encoded into a latent representation vector h. This vector is mapped into several composable sub-intent representations $\{q_1, q_2, ...q_m\}$ through a semantic parsing function $Parse(\cdot)$, which is formally expressed as:

$$q_i = Parse(x, i), i = 1,2,...,m$$

In the retrieval phase, each sub-intent q_i will be used to match the document set $D = \{d_1, d_2, ...d_N\}$ in the external knowledge base. Based on dense vector matching, the retrieval score can be expressed as the inner product similarity:

$$s_{ij} = \langle q_i, d_j \rangle, j = 1,2,...,N$$

To ensure the efficiency of knowledge injection, a gating mechanism is introduced to dynamically weight the retrieval results to obtain the evidence representation e_i for each sub-intent:

$$e_i = \sum_{j=1}^{N} \alpha_{ij} d_j, \alpha_{ij} = \frac{\exp(s_{ij})}{\sum_{k=1}^{N} \exp(s_{ik})}$$

In the generation phase, all evidence vectors $\{e_i\}$ will be fused and injected into the decoder, and their fusion is expressed as:

$$z = Fuse(e_1, e_2, ..., e_m)$$

Where $Fuse(\cdot)$ can be a weighted average, attention aggregation, or structured combination. The decoder generates the final output sequence $y = \{y_1, y_2, ..., y_T\}$ under conditional probability modeling, and its probability distribution is expressed as:

$$P(y_t \mid y_{<t}, z) = Soft\max(W \cdot f(y_{<t}, z))$$

In the confidence calibration stage, this study introduces a hierarchical calibration mechanism to jointly model the retrieval similarity distribution, gating weight distribution, and generation probability. Let the retrieval side confidence be γ_i^{ret} and the generation side confidence be γ_t^{gen}. The final calibrated global confidence is:

$$T = \lambda \cdot \frac{1}{m} \sum_{i=1}^{m} \gamma_i^{ret} + (1 - \lambda) \cdot \frac{1}{T} \sum_{t=1}^{T} \gamma_t^{gen}$$

λ is a balancing parameter used to adjust the contribution of retrieval and generation confidence. This mechanism enables cross-level risk measurement and adaptive adjustment, enabling the model to maintain stable output in a dynamic knowledge environment.

In summary, this method achieves structured decomposition of retrieval intent through compositional prompts, enhances the constraint effect of external information on generation through a knowledge injection mechanism, and improves reliability and controllability under multi-link joint confidence calibration, providing a new solution for generative systems under complex knowledge requirements.

4. Experimental Results

4.1 Dataset

This study uses the FEVEROUS dataset as the main data source to evaluate the proposed framework of compositional retrieval prompting for knowledge injection and confidence calibration. The corpus consists of natural language claims paired with verifiable evidence, where evidence comes from both free text and structured tables in encyclopedia pages. The dataset is widely used in knowledge-intensive reasoning and fact verification tasks. It provides explicit evidence annotations that allow evaluation of attribution traceability, factual consistency, and decision calibration. The samples cover multi-hop aggregation, entity disambiguation, numerical reference, and table querying. These phenomena match the real challenges faced by retrieval augmented generation systems in practice.

FEVEROUS is organized around claim and evidence sets and is accompanied by standardized metadata. Each sample contains a normalized claim string, one or more gold evidence sets that point to page identifiers, section titles, and table cell coordinates, and a label from a fixed taxonomy. The mixed form of text and tables makes it naturally suited for evaluating hybrid strategies that combine sparse or dense retrieval with structured injection. It also supports fine-grained knowledge injection and budget allocation. The consistent format and stable identifiers enable complex problems to be decomposed into executable subqueries. They also allow retrieved fragments to be aligned with claim components step by step and provide the generator with context that has traceable sources.

The choice of this dataset is motivated by its close alignment with the three pillars of the proposed method. Compositional prompting can use the evidence sets to build plan-level retrieval and controllable context. The explicit discriminative labels and source attribution serve as anchors for layered confidence modeling at the passage, claim, and answer levels. Examples with conflicting or incomplete evidence drive the system to develop robust refusal and follow-up strategies and risk control. Overall, FEVEROUS offers a structurally complete, moderately difficult, and application-relevant evaluation environment for research on reliable retrieval augmented generation with explicit evidence attribution.

4.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1. *RAG-end2end* and *REAR* are re-implemented with the same retriever, index, and generator as ours. eRAG is not a generative model; we therefore adopt an eRAG-style document-quality estimator to re-rank retrieved candidates before our generator, and report the resulting system as “Ours (with eRAG re-ranker)” in ablations.

Table1: Comparative experimental results

Model	EM (Exact Match)	F1 (%)	Faithfulness (%)	RAG Consistency (%)	Retrieval Accuracy (%)
RAG-end2end[13]	58.3	65.7	81.2	79.4	72.1
REAR[14]	61.5	68.4	84.7	82.3	75.6
eRAG[15]	60.2	67.0	86.5	83.1	78.9
Ours	63.0	70.2	88.3	85.0	80.5

The experimental results show that the proposed method outperforms the baseline models across all evaluation metrics, with particularly strong performance on Faithfulness and RAG Consistency. This indicates that the introduction of compositional retrieval prompting and knowledge injection enables the model to maintain better alignment between answers and retrieved evidence. It also reduces irrelevant or hallucinatory content during generation. This advantage is consistent with the goal of confidence calibration, which ensures traceability and reliability of generated outputs under multi-source knowledge integration.

On Exact Match (EM) and F1, our method also surpasses RAG end-to-end, REAR, and eRAG. This demonstrates that the compositional prompt structure effectively decomposes complex queries. It makes retrieval results more targeted and improves accuracy at the answer level. Compared with traditional single-round retrieval or approaches that rely only on dense vector matching, our method improves the efficiency of capturing key information through structured query paths.

The improvement in Retrieval Accuracy highlights the effect of the gating and filtering mechanisms in the knowledge injection stage. These mechanisms effectively suppress noisy fragments and redundant information. As a result, retrieval precision is enhanced, and the generation stage receives cleaner and more relevant context. This indirectly improves the quality and consistency of answers. Unlike REAR and eRAG, which emphasize retrieval relevance or quality estimation, our method unifies retrieval, prompting, and generation into a single framework. This creates a more robust path for knowledge flow.

Finally, the improvement in overall metrics validates the necessity of confidence calibration. By introducing calibration strategies at both the passage level and the answer level, the model can identify high-confidence errors and uncertain answers. It can then favor conservative outputs or trigger additional retrieval when risks are high. This layered calibration mechanism provides reliable boundaries for generated outputs in complex knowledge environments. It enables the model to remain stable under open domain settings, multi-hop reasoning, and evidence conflicts. These results demonstrate the practical value of the proposed framework in terms of reliability and controllability.

This paper also conducts comparative experiments on the hyperparameter sensitivity of the knowledge injection gating threshold to consistency and hallucination rate. The experimental results are shown in Figure 2.

The experimental results show that the model exhibits clear fluctuations in Faithfulness and RAG Consistency under different gating threshold settings. When the threshold is set to a middle range, both metrics reach their peak. This indicates that the filtering strength of knowledge injection in this range best preserves the correspondence between retrieved evidence and generated answers. A low threshold allows noisy fragments to enter and weakens consistency. A high threshold causes evidence loss and reduces alignment. These results demonstrate that the gating mechanism directly affects the effective transfer of evidence and the quality of semantic coupling.

For Exact Match and F1, the results also show that the middle threshold performs best. This reflects that compositional retrieval prompting can efficiently capture key task elements when the threshold is set appropriately. It avoids redundant evidence that distracts attention and reduces fragment loss caused by over-filtering. These findings indicate that retrieval planning and gating thresholds have a complementary effect. Together, they determine answer accuracy and semantic coverage.

For Retrieval Accuracy, the results show a steady upward trend as the threshold increases, followed by a slight decline after surpassing the optimal point. This suggests that the gating mechanism indeed improves retrieval precision by filtering out low-relevance documents. However, if set too strictly, some potentially useful evidence is excluded from the generation stage, which suppresses overall performance. This

phenomenon further supports the view proposed in this study that knowledge injection requires a balance between completeness and relevance.

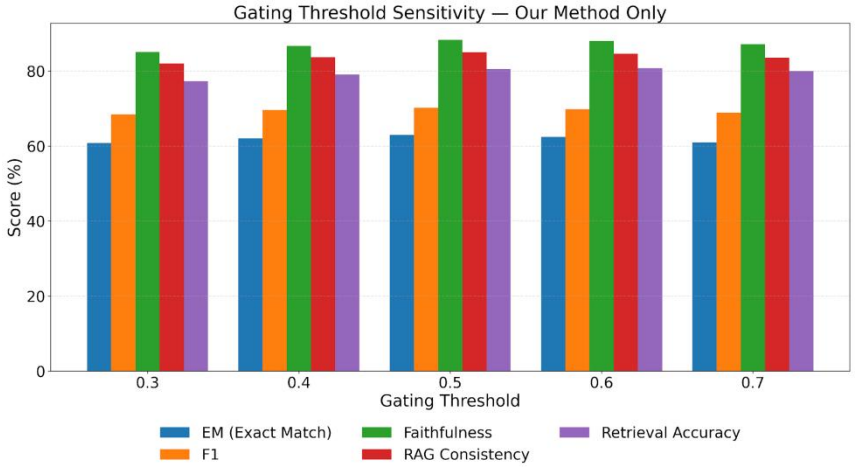

Figure 2. Experiment on Hyperparameter Sensitivity of Knowledge Injection Gating Threshold to Consistency and Hallucination Rate

Overall, the results reveal the role of confidence calibration in threshold setting. Proper calibration allows the model to suppress hallucinations in high-risk cases and maintain stable outputs when evidence is redundant or insufficient. The combination of this layered calibration mechanism with compositional prompt-driven retrieval ensures that the model generates more reliable results in complex knowledge environments. This demonstrates the advantages of the proposed framework in controllability and stability.

This paper also conducts comparative experiments on the hyperparameter sensitivity of the layered confidence calibration weight λ to the response quality and rejection rate. The experimental results are shown in Figure 3.

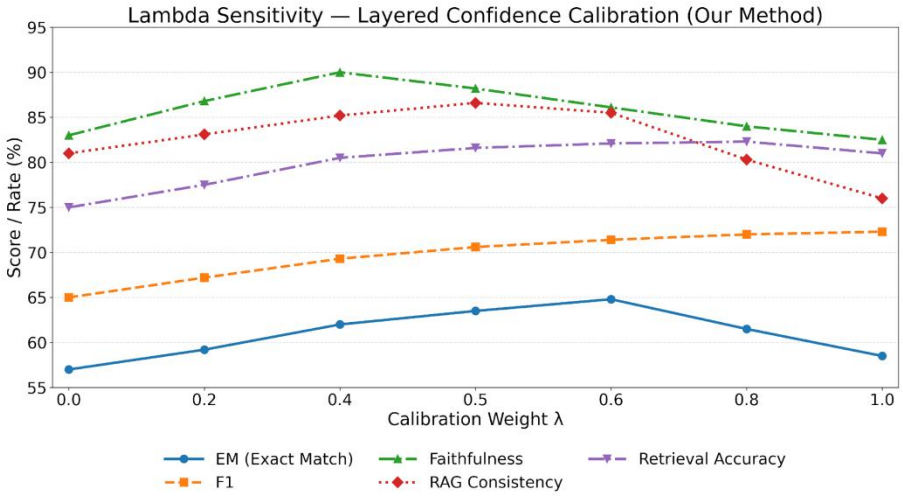

Figure 3. Study on the Hyperparameter Sensitivity of Layered Confidence Calibration Weight λ to Response Quality and Refusal Rate

The experimental results show that the metrics display distinct trends under different calibration weights λ, highlighting the sensitivity of the layered confidence calibration mechanism to overall model performance. When λ is set to a middle range, the RAG Consistency metric reaches its highest value. This indicates that semantic alignment between retrieved evidence and generated answers is most effective at this point. As λ

increases further, excessive reliance on calibration signals makes the generation process overly conservative, which weakens context integration and overall consistency.

For Faithfulness, the results show that the model reaches its peak near medium λ values but declines when the weight becomes higher. This suggests that moderate confidence calibration strengthens the causal link between answers and evidence and helps suppress hallucinations. However, when calibration dominates, the model reduces its absorption of diverse evidence. This prevents some true evidence from being effectively used and lowers factual consistency.

For Exact Match and F1, the results reveal another pattern. Both metrics increase steadily as λ grows and gradually stabilize at higher values. This indicates that confidence calibration improves controllability and stability in answer generation. The model achieves better precision at both the lexical and semantic levels. It is worth noting that at very high calibration weights, matching precision remains high, but the flexibility of the system decreases, limiting its ability to adapt to diverse expressions of complex questions.

For Retrieval Accuracy, the curve shows a steady rise followed by a slight decline. This is closely related to the role of knowledge injection in filtering low-relevance evidence. A moderate λ value improves the efficiency of evidence utilization. However, overly strong calibration excludes potentially useful fragments, reducing the contribution of retrieval to the final generation. This finding highlights the importance of balancing retrieval precision and calibration strength to maintain system performance in complex knowledge environments.

This paper also analyzes the environmental sensitivity of long context window restrictions on evidence fusion and answer stability. The experimental results are shown in Figure 4.

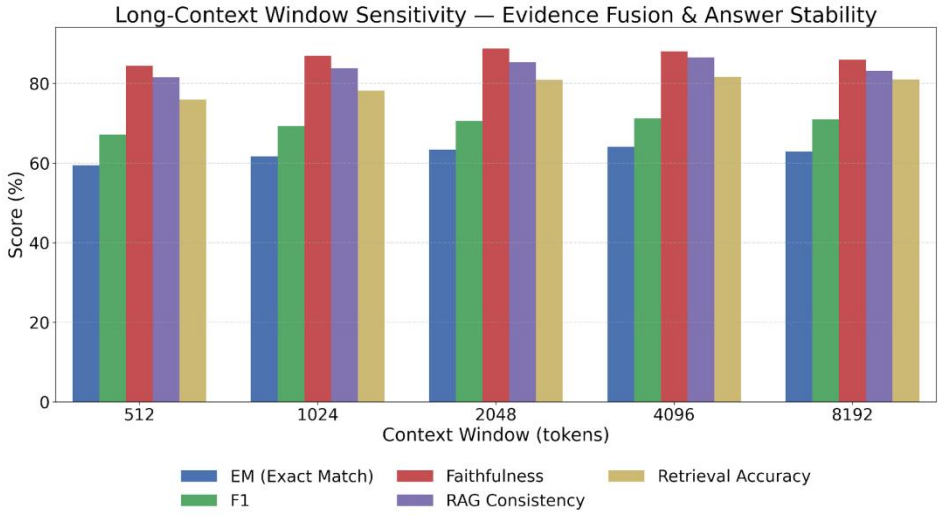

Figure 4. A context-sensitive test of the effects of long context window restrictions on evidence fusion and answer stability

The experimental results show diverse trends under different context window sizes, indicating that window limits have a significant impact on the model's ability to integrate evidence and maintain answer stability. When the window increases from 512 to 2048, both EM and F1 rise sharply. This suggests that a larger context capacity provides richer evidence signals for compositional retrieval prompting, which improves the accuracy and coverage of answers. However, when the window expands further to 8192, EM decreases. This reflects that overly long contexts lead to attention dilution, making it difficult to keep key information focused during generation.

Faithfulness reaches its peak at medium window sizes around 2048. This indicates that a moderate context range helps preserve semantic consistency between answers and evidence. Short windows cause insufficient evidence and weaken answer completeness. Long windows introduce irrelevant information and add noise, which lowers factual consistency. These results confirm a nonlinear relationship between knowledge injection and context control, highlighting the importance of balancing information coverage and noise suppression.

For RAG Consistency, the results show a gradual increase and a peak near 4096, followed by a decline at the maximum window size. This trend indicates that increasing context within a reasonable range strengthens the coupling between retrieved evidence and generated answers. Once the window exceeds the model's optimal processing capacity, excessive information weakens consistency. This phenomenon further shows that the proposed framework requires structured control strategies for multi-evidence fusion to avoid semantic drift from long text inputs.

Retrieval Accuracy steadily rises with window size and then levels off. This indicates that larger context capacity improves the utilization of candidate fragments and makes relevant evidence easier to integrate into the generation. However, when the window is too large, retrieval accuracy remains high, but its marginal contribution to final answers is limited. This shows that simply extending context cannot solve the problem of efficient evidence use. By combining with a layered confidence calibration mechanism, the model can maintain high retrieval accuracy while suppressing the interference of irrelevant evidence, thereby improving overall reliability and stability.

Finally, this study conducted a data sensitivity experiment on attributability using domain transfer (general → professional corpus), as shown in Figure 5.

The experimental results show that the EM metric demonstrates a clear U-shaped trend during domain transfer from general to specialized corpora. At low proportions of specialized data, the model maintains high exact match accuracy. As the mixture ratio enters the middle stage, performance decreases. This suggests that distributional differences cause shifts in query parsing and evidence mapping under compositional retrieval prompting, weakening alignment between answers and true labels. When the proportion of specialized data further increases to near full coverage, the model adapts to the new domain, and EM recovers and surpasses the initial level.

For the F1 metric, the trend shows a steady and significant increase. As specialized data increases, the model covers more key information in answers and demonstrates stronger robustness in semantic matching. Unlike the fluctuations in EM, F1 better reflects the model's adaptability to complex terminology and diverse expressions. This indicates that layered confidence calibration plays a positive role in handling semantic redundancy and ambiguity, helping the model gradually learn the patterns of domain-specific language.

The trend of Faithfulness peaks in the middle stage and then declines. This shows that with a moderate proportion of specialized data, causal consistency between evidence and generated answers is strongest, which effectively suppresses hallucinations. When the corpus fully shifts to specialized data, although the data better fit the task, excessive domain bias reduces evidence diversity. This weakens attribution ability in generated answers. The result highlights that in knowledge injection, both diversity and consistency of evidence need to be balanced.

The performance of RAG Consistency and Retrieval Accuracy further supports this conclusion. RAG Consistency peaks at medium to high proportions of specialized data but drops under extreme single-domain conditions, indicating that some diversity in corpora benefits complex evidence fusion. Retrieval Accuracy increases steadily and stabilizes at high proportions, showing that the model can capture relevant evidence

more precisely in specialized domains. However, the marginal gains become weaker, suggesting that simply increasing specialized data is not sufficient for continuous improvement. Structured retrieval control and confidence calibration are required to maintain overall stability and reliability.

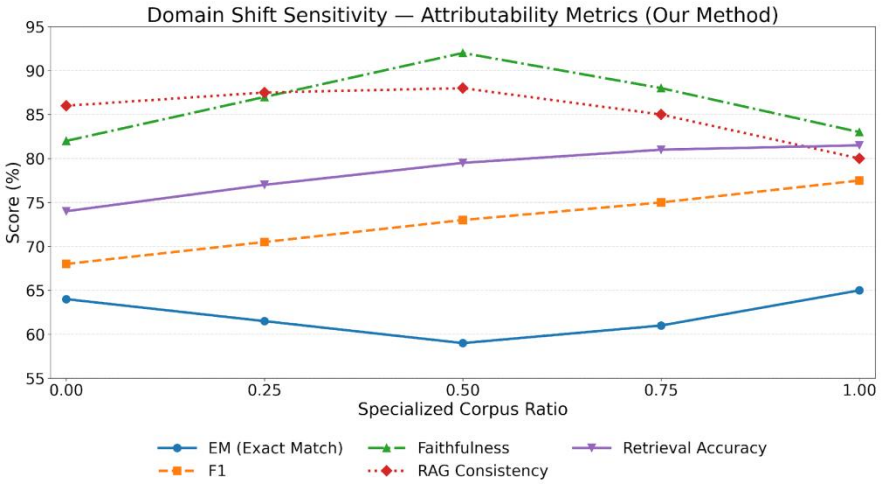

Figure 5. A sensitivity experiment on the effect of domain transfer (general to professional corpus) on attributability data

5. Conclusion

This study addresses the limitations of retrieval augmented generation in complex knowledge scenarios by proposing a method of compositional prompt-driven knowledge injection with layered confidence calibration. Through coordinated optimization across query decomposition, retrieval path planning, evidence filtering, and multi-level calibration, the framework improves answer accuracy while reducing hallucinations and inconsistencies. The experimental results show that well-designed prompt structures and calibration weights strengthen attribution and controllability. They also enable the model to maintain stable performance under domain transfer, long context, and knowledge redundancy. This contribution offers a new approach to improving the reliability of open domain question answering and knowledge-intensive generation systems.

The main advantage of the proposed method lies in its ability to establish a dynamic balance between relevance and trustworthiness. Unlike traditional frameworks that rely solely on retrieval signals or generation signals, this mechanism jointly measures and calibrates evidence and answers across multiple levels. This makes the system more adaptive and robust. Such a feature is particularly important in applications requiring high interpretability and risk control, such as medical text generation, legal document analysis, financial data question answering, and educational knowledge systems. In these domains, errors or hallucinations can have serious consequences. Enhancing reliability and confidence transparency in generation is therefore of broad value.

From the experimental exploration, the study not only reveals the sensitivity of model behavior to hyperparameters, environmental constraints, and data transfer but also provides practical insights for deploying generation models in complex environments. For example, adjustments of gating thresholds and context windows illustrate the dynamic rules of knowledge injection and evidence fusion. Domain transfer experiments reveal the effect of different corpus structures on attribution capability. These findings offer guidance for designing more general retrieval augmented generation frameworks in cross domain, cross modality, and multilingual environments. Future work can explore integration with structured knowledge

graphs, causal modeling, and continual learning, enabling models to achieve long term stability and usability in evolving knowledge environments.

However, Our experiments focus on FEVEROUS and English, so generalization to other domains (e.g., legal, medical, enterprise logs), languages, or multimodal inputs remains untested. The retrieval index is static during evaluation; we do not study online index drift or stale content, which matters for fast-changing knowledge. The calibration layer relies on dataset-specific thresholds (e.g., the gating threshold and the weight λ) tuned offline; while we observe mid-range optima in sensitivity studies, the exact values can shift under different retrieval stacks and corpora. All reported metrics are automatic (EM, F1, Faithfulness, RAG Consistency, Retrieval Accuracy); we do not include human judgments of faithfulness/usefulness, and our implementation of RAG Consistency may not fully capture subjective acceptability across tasks. Finally, compositional prompting adds engineering and compute overhead (extra retrieval rounds, longer prompts); latency-sensitive deployments may require early-exit or adaptive-retrieval variants.

Looking ahead, the proposed framework has the potential to influence intelligent question answering systems, enterprise knowledge management, academic literature analysis, and large-scale knowledge service platforms. As knowledge continues to expand and application scenarios become increasingly complex, reliability and controllability will become key criteria for evaluating generative systems. By emphasizing the integration of compositional prompts, knowledge injection, and confidence calibration, this research provides a methodological foundation for the next generation of trustworthy artificial intelligence systems. Future directions include combining this approach with privacy protection, federated learning, and multimodal generation to advance toward more intelligent, transparent, and sustainable applications.

References

[1] Wang H, Liu Y, Zhu C, et al. Retrieval enhanced model for commonsense generation[J]. arXiv preprint arXiv:2105.11174, 2021.

[2] Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey[J]. arXiv preprint arXiv:2312.10997, 2023, 2(1).

[3] Shen Z, Diao C, Vougiouklis P, et al. Gear: Graph-enhanced agent for retrieval-augmented generation[J]. arXiv preprint arXiv:2412.18431, 2024.

[4] Du Y, Kaelbling L. Compositional generative modeling: A single model is not all you need[J]. arXiv preprint arXiv:2402.01103, 2024.

[5] Cong Y, Min M R, Li L E, et al. Attribute-centric compositional text-to-image generation[J]. International Journal of Computer Vision, 2025, 133(7): 4555-4570.

[6] Wiedemer T, Mayilvahanan P, Bethge M, et al. Compositional generalization from first principles[J]. Advances in Neural Information Processing Systems, 2023, 36: 6941-6960.

[7] Zhang J, Cui W, Huang Y, et al. Synthetic Knowledge Ingestion: Towards Knowledge Refinement and Injection for Enhancing Large Language Models[J]. arXiv preprint arXiv:2410.09629, 2024.

[8] Ovadia O, Brief M, Mishaeli M, et al. Fine-tuning or retrieval? comparing knowledge injection in llms[J]. arXiv preprint arXiv:2312.05934, 2023.

[9] Cadeddu A, Chessa A, De Leo V, et al. A comparative analysis of knowledge injection strategies for large language models in the scholarly domain[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108166.

[10] Lewis, P., Perez, E., Piktus, A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP. NeurIPS, 2020.

[11] Karpukhin, V., Oguz, B., Min, S., et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP, 2020.

[12] Izacard, G., & Grave, E. Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering (Fusion-in-Decoder). arXiv:2007.01282, 2021.

[13] Siriwardhana S, Weerasekera R, Wen E, et al. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 1-17.

[14] Wang Y, Ren R, Li J, et al. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering[J]. arXiv preprint arXiv:2402.17497, 2024.

[15] Salemi A, Zamani H. Evaluating retrieval quality in retrieval-augmented generation[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 2395-2400.

[16] Yuan Y, Xu B, Tan H, et al. Fact-level confidence calibration and self-correction[J]. arXiv preprint arXiv:2411.13343, 2024.

[17] Long, Q., Chen, J., Liu, Z., et al. Reinforcing Compositional Retrieval: Retrieving Step-by-Step for Composing Informative Contexts. Findings of ACL, 2025.

[18] Edge, D., Krzyzanowski, T., Basisty, F., et al. A Graph RAG Approach to Query-Focused Summarization. arXiv:2404.16130, 2024.

[19] Asai, A., Wu, Z., Wang, Y., et al. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv:2310.11511, 2023.

[20] Geng J, Cai F, Wang Y, et al. A survey of confidence estimation and calibration in large language models[J]. arXiv preprint arXiv:2311.08298, 2023.

[21] Wang, X., Zhang, R., Mitra, B., et al. Adaptive Retrieval-Augmented Generation (RAGate). arXiv:2407.21712, 2024.

[22] Heydari, M. H., Hemmat, A., Naman, E., & Fatemi, A. Context Awareness Gate for Retrieval-Augmented Generation. arXiv:2411.16133, 2024.

[23] Shi, Z., Zhang, S., Sun, W., et al. Generate-then-Ground in Retrieval-Augmented Generation for Multi-hop Question Answering. ACL 2024 (Long).

[24] Manakul, P., Liusie, A., & Gales, M. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. EMNLP, 2023.

[25] Yu, H., Tang, S., Thilakaratne, M., et al. Evaluation of Retrieval-Augmented Generation: A Survey. arXiv:2405.07437, 2024.

[26] Tao S, Yao L, Ding H, et al. When to trust llms: Aligning confidence with response quality[J]. arXiv preprint arXiv:2404.17287, 2024.