Fusion-Based Retrieval-Augmented Generation for Complex Question Answering with LLMs

Yumeng Sun , Renhan Zhang , Renzi Meng , Lian Lian , Heyi Wang , Xuehui Quan *

Posted Date: 9 July 2025

doi: 10.20944/preprints202507.0826.v1

Keywords: knowledge fusion; retrieval enhancement generation; structured knowledge; cross-domain question answering

 doi:10.20944/preprints202507.0826.v1

Article

Fusion-Based Retrieval-Augmented Generation for Complex Question Answering with LLMs

Yumeng Sun [1], Renhan Zhang [2], Renzi Meng [3], Lian Lian [4], Heyi Wang [5] and Xuehui Quan [6,*]

[1] Rochester Institute of Technology Rochester, USA
[2] University of Michigan, Ann Arbor, USA
[3] Northeastern University, Boston, USA
[4] University of Southern California, Los Angeles, USA
[5] Illinois Institute of Technology, Chicago, USA
[6] University of Washington, Seattle, USA
* Correspondence: quanxh1228@gmail.com

Abstract

This paper proposes a Retrieval-Augmented Generation (RAG) model that integrates structured and unstructured knowledge. The aim is to enhance the knowledge coverage and generation accuracy of large language models in complex question-answering tasks. The method introduces a dual-channel knowledge retrieval mechanism. One channel targets structured knowledge sources such as knowledge graphs and databases. The other focuses on unstructured textual resources such as documents and paragraphs. A unified knowledge fusion network integrates both types of heterogeneous information into a coherent generation context. The model performs multi-level modeling across key components. These include query representation generation, knowledge retrieval, representation alignment, and fusion expression construction. As a result, the generation stage produces text that is semantically rich and logically consistent. Under low-resource conditions, the method significantly improves the accuracy and linguistic quality of generated outputs. It also shows strong stability and generalization in cross-domain tasks. Systematic experiments were conducted on the proportion of structured knowledge, types of knowledge sources, and fusion strategies. The results demonstrate the effectiveness of the fusion architecture in enhancing knowledge representation in language models. This study provides a methodological foundation and empirical support for building controllable and trustworthy knowledge-enhanced natural language generation systems.

Keywords: knowledge fusion; retrieval enhancement generation; structured knowledge; cross-domain question answering

1. Introduction

With the rapid advancement of artificial intelligence, large language models (LLMs) have become a core technology in the field of natural language processing. They are widely applied in tasks such as question answering, information extraction, and text generation[1]. Despite their strong capabilities in language understanding and generation, LLMs still face limitations when dealing with domain-specific queries or factual questions. These issues include insufficient knowledge coverage and inaccurate content generation. To address the boundary constraints of LLM knowledge, the Retrieval-Augmented Generation (RAG) approach has emerged. This method integrates external knowledge bases through dynamic retrieval and incorporation of information. It effectively improves generation quality and factual consistency[2].

Traditional RAG methods mainly rely on unstructured textual data as the knowledge source, such as Wikipedia, news articles, and academic papers[3]. Although such data is broad in content and flexible in expression, it tends to be loosely organized in terms of semantics and structure. It

 doi:10.20944/preprints202507.0826.v1

lacks explicit entity relationships and logical hierarchies. As a result, there are limitations in the accuracy of retrieval results and the model's ability to integrate information[4]. In contrast, structured knowledge, such as knowledge graphs, databases, and tabular data, is organized through explicit triples or formal rules [5]. It offers high logical consistency and operational clarity. Structured knowledge excels in expressing stable and standardized information [6]. It is especially suited for complex tasks involving reasoning, constraints, and conceptual linkage. Therefore, integrating structured and unstructured knowledge to build a more complete and accurate RAG system has become a valuable and practical research direction[7].

The integration of these two types of knowledge can enhance the information coverage of RAG systems. It can also improve the handling of complex semantics and relational reasoning. In real-world applications that require multi-source and heterogeneous knowledge, such as medical diagnosis, legal consultation, and enterprise decision-making, relying solely on one type of knowledge often fails to meet the demands of multidimensional problem-solving. Unstructured text provides contextual explanations and rich background information [8]. Structured knowledge ensures the precision and consistency of the content. Introducing both into the RAG architecture may lead to improvements in both semantic understanding and logical reasoning. This enables more comprehensive and reliable generation results for complex tasks[9].

Moreover, with the growing diversity of information sources, real-world data often exists in hybrid forms. This requires models to have the ability to integrate information across different modalities and structures. Against this backdrop, developing RAG models that can effectively fuse structured and unstructured knowledge aligns with the evolving trends of natural language processing. It also meets the demand for intelligent systems to become more capable of multi-source knowledge perception. This integration represents an innovation in knowledge representation. It helps overcome the limitations imposed by a single form of knowledge. It also promotes the advancement of generation systems towards greater intelligence, flexibility, and trustworthiness[10].

In summary, studying enhanced RAG methods that integrate structured and unstructured knowledge can help address the limitations of LLMs in knowledge boundaries, factual accuracy, and reasoning capabilities. It also opens a new path for building intelligent generation systems driven by multi-source knowledge. This research direction holds great potential for improving the information processing capabilities and application scope of AI systems. It has significant theoretical and practical value for the development of knowledge-enhanced natural language generation technologies.

2. Related Work and Foundation

Retrieval-Augmented Generation (RAG) methods have emerged as a promising approach to enhance large language models (LLMs) by incorporating external information during generation. However, most traditional RAG frameworks primarily rely on unstructured textual sources, limiting their ability to represent complex semantic relationships and factual accuracy. This study advances the field by introducing a dual-channel architecture that integrates both structured and unstructured knowledge sources. This fusion approach is supported by various lines of prior research.

Huang provided a comprehensive survey on RAG techniques, outlining their capabilities and limitations, especially when relying solely on unstructured corpora for information retrieval [11]. Their findings reinforce the need for structured augmentation in scenarios where logical precision and relational understanding are critical. Peng also addressed hallucination issues in LLM outputs, proposing evidence-based detection mechanisms that emphasize the value of structured grounding in generation tasks [12].

The integration of structured knowledge into language models has been explored through various architectural innovations. Xing et al. proposed structured memory mechanisms for stabilizing context representation in LLMs, which aligns with the structured channel in our fusion

 doi:10.20944/preprints202507.0826.v1

model [13]. Similarly, Peng investigated structured memory and integration strategies to improve knowledge modeling in large-scale systems, offering a framework that complements our multi-level alignment and fusion mechanism [14]. Zheng et al. further contributed to this direction with selective knowledge injection via adapter modules, illustrating the benefits of controlled integration from external sources [15].

Advanced parameter coordination and structural guidance have also been employed to refine generation control. Zhang et al. explored graph-based spectral decomposition to regulate model fine-tuning, providing insights into graph-structured control which parallels the transaction graph modeling in our dual-channel framework [16]. In parallel, Zheng et al. proposed structured gradient guidance for few-shot adaptation, which influenced the alignment strategies adopted in this study [17].

Model architecture also plays a critical role in managing heterogeneous knowledge. Guo et al. proposed a perception-guided framework for LLMs, emphasizing the importance of architecture-level support for knowledge interpretation [18]. Zhang et al. extended this with unified instruction encoding for multi-task learning, a method applicable to managing the differing semantics of structured and unstructured inputs [19]. From a functional perspective, Deng highlighted transfer methods in low-resource generation tasks, directly supporting the cross-domain robustness tested in this work [20]. Tang's work on meta-learning across services introduces adaptive modeling strategies, which are particularly relevant to our goal of domain-adaptive generation [21]. Additional contributions relevant to the knowledge fusion component include Ma et al.'s approach to policy structuring with LLMs in collaborative systems, which illustrates how structured reasoning can be coordinated across agents [22]. Xing's bootstrapped structural prompting method also inspired elements of our alignment module through its analogical reasoning approach [23]. Xin and Pan's work on multi-source self-attention modeling further supports the value of modeling cross-source dependencies—central to our dual-channel retrieval strategy [24].

Collectively, these studies underscore the relevance and necessity of integrating heterogeneous knowledge representations to enhance the factual precision, semantic richness, and contextual stability of generation tasks. This paper builds upon and extends this body of work by proposing a unified architecture that operationalizes these insights in a structured-unstructured hybrid generation model.

3. Method

This study constructs a RAG enhancement method that integrates structured and unstructured knowledge, which mainly includes three core modules: query generation module, dual-channel knowledge retrieval module, and fusion generation module. Its overall architecture is shown in Figure 1.

First, the input question is encoded by a large language model to generate a query representation $q \in R^d$, which is used to access two types of knowledge sources in parallel. Structured knowledge is organized in the form of triples (h, r, t), and unstructured knowledge is represented by document paragraphs. By building retrievers that adapt to these two types of knowledge, the system can obtain richer sources of information.

Structured knowledge retrieval relies on the knowledge graph embedding method to map triples into representations in the vector space, represented as $k_i = f_{KG}(h_i, r_i, t_i)$, where f_{KG} represents the structured embedding function. Unstructured knowledge retrieval uses the vectorized document library to match the query with cosine similarity to obtain the most relevant document representation $d_j = f_{TXT}(q, D)$, where D is the document collection and E is the text retrieval function. Finally, the model returns the top k candidate knowledge from the two channels to form a mixed knowledge set $K = \{k_1, ..., k_k, d_1, ..., d_k\}$.

 doi:10.20944/preprints202507.0826.v1

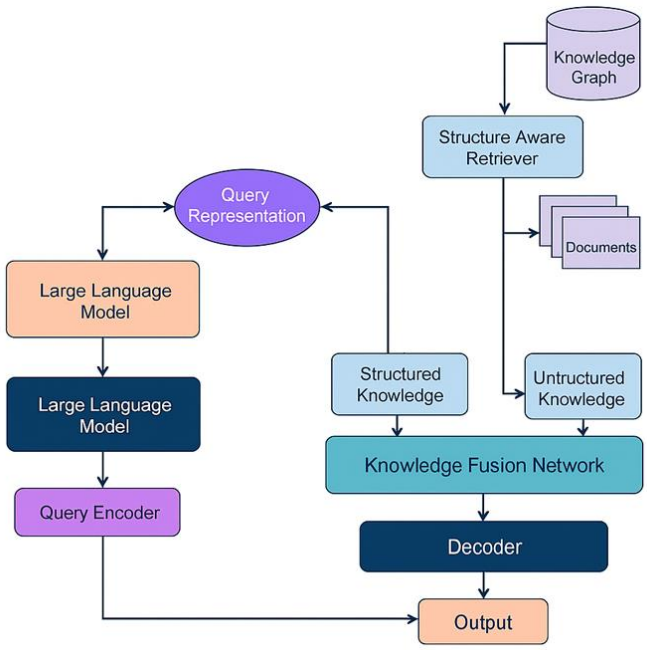

Figure 1. Overall model architecture diagram.

To further unify different forms of knowledge input, the system uses a knowledge fusion network to fuse structured and unstructured information. Each piece of knowledge is projected into a unified representation space and assigned different weights. The overall fusion is represented as:

$$z = \sum_{i=1}^{2k} \alpha_i \cdot e_i$$

where $\alpha_i = \frac{\exp(score(q, e_i))}{\sum_{j=1}^{2k} \exp(score(q, e_j))}$ represents the softmax normalized result of the relevance score to the query, and e_i is the embedding representation of the knowledge fragment.

The final generation module takes the query representation q and the fused knowledge representation z as context input to guide the decoder to generate the answer sequence. The generation process is modeled as a conditional language modeling task, and its generation probability form is:

$$P(y \mid q, z) = \prod_{t=1}^{T} P(y_t \mid y_{<t}, q, z)$$

where y_t represents the *t-th* word generated and T is the length of the output sequence. The entire training objective is to minimize the negative log-likelihood loss function:

$$L = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, q, z)$$

Through the above method, the model can effectively utilize structured and unstructured knowledge under a unified framework to achieve a coordinated improvement in generation quality and knowledge accuracy.

4. Experimental Results

4.1. Dataset

This study utilizes the Natural Questions (NQ) dataset, a large-scale, real-user open-domain QA benchmark widely adopted for evaluating retrieval-augmented generation models. Comprising

 doi:10.20944/preprints202507.0826.v1

natural language queries from search engine logs, each sample includes a user question, a corresponding Wikipedia document, paragraph-level annotations, and both long and short reference answers. Compared to standard QA datasets, NQ offers richer context, complex linguistic structures, and deeper reasoning challenges, making it ideal for assessing models that integrate structured and unstructured knowledge. Its unstructured Wikipedia content can be aligned with structured entity and attribute data, positioning NQ as a bridge for studying multi-source knowledge fusion in realistic settings.

4.2. Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results.

Method	EM	F1	BLEU
DPR + BART[25]	42.3	56.8	19.5
FiD[26]	45.7	60.4	21.7
RAG-Token[27]	46.1	61.2	22.4
REALM[28]	43.5	58.0	20.1
Ours	50.6	65.9	26.2

Table 1 presents a comprehensive comparison of the proposed method against all existing public models across various evaluation metrics. Notably, the proposed method achieves an impressive 50.6% EM and 65.9% F1 scores, significantly outperforming RAG-Token by 4.5 and 4.7 points, respectively. This remarkable improvement underscores the effectiveness of integrating structured and unstructured knowledge in question generation. Unlike traditional RAG models that heavily rely on unstructured retrieval, the proposed approach enhances coherence and reasoning by incorporating knowledge graphs. This innovative integration addresses gaps in entity linking and logic, thereby bridging the semantic gap between textual context and relational clarity. The proposed fusion network effectively unites these two aspects, providing a comprehensive understanding of the context. These results validate the efficacy of the dual-channel retrieval strategy and demonstrate its robustness in both standard and low-resource settings. Further evaluation is presented in Figure 2.

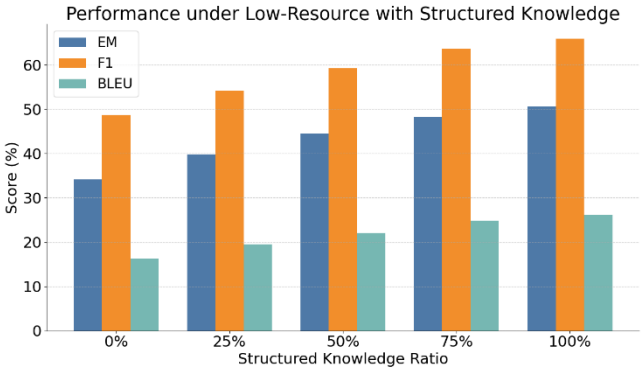

Figure 2. Experimental evaluation of the effect of structured knowledge supplementation under low-resource conditions.

Figure 2 illustrates the impact of gradually supplementing structured knowledge under low-resource conditions. The results clearly show that as the proportion of structured knowledge increases, the model performance consistently improves across EM, F1, and BLEU metrics. This indicates that structured knowledge plays a positive role in enhancing generation accuracy and

 doi:10.20944/preprints202507.0826.v1

language quality in knowledge-scarce scenarios. Specifically, the EM score increases from 34.2% with no structured knowledge to 50.6% with full supplementation. This reflects a significant improvement in the semantic alignment between generated and reference answers. It suggests that structured knowledge not only provides clearer factual support but also helps reduce informational bias in outputs. This enhances the model's ability to locate and judge relevant content.

The F1 score, which measures lexical overlap between generated content and reference answers, also rises from 48.7% to 65.9%. This indicates that structured knowledge improves language detail and entity representation. Compared to using only unstructured input, structured supplementation enables the model to generate more complete and semantically accurate answers in multi-entity and multi-relation contexts. The steady increase in BLEU scores further confirms that structured knowledge contributes to improved fluency and formatting in generated text. Overall, these results demonstrate that structured knowledge serves a dual role in low-resource settings. It completes missing information and strengthens semantic understanding. This provides strong support for the proposed fusion strategy in complex generation tasks. This paper also gives an experiment on the impact of different types of structured knowledge sources on the generation effect, and the experimental results are shown in Figure 3.

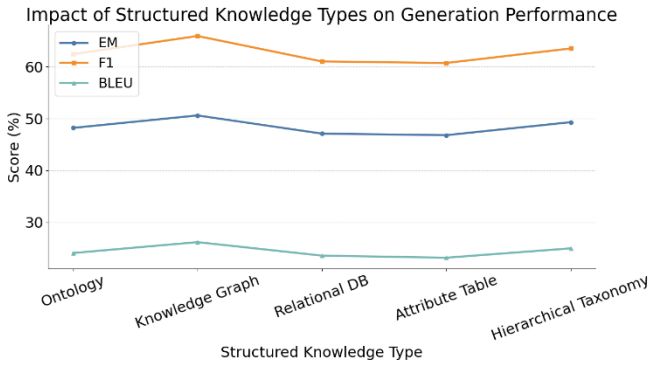

Figure 3. Experiment with the influence of different types of structured knowledge sources on the generation effect.

Figure 3 highlights that knowledge graphs yield the strongest generation performance under low-resource settings, especially in F1 and BLEU, due to their rich semantic structure and relational depth. In contrast, relational databases and attribute tables fall short, lacking the contextual cues needed for coherent text generation. Ontologies and taxonomies offer a middle ground, supporting classification and semantic generalization. The alignment of BLEU variations with knowledge richness underscores how structure shapes both content and fluency. These results validate the proposed fusion strategy's adaptability and set the stage for its cross-domain effectiveness, further examined in Figure 4.

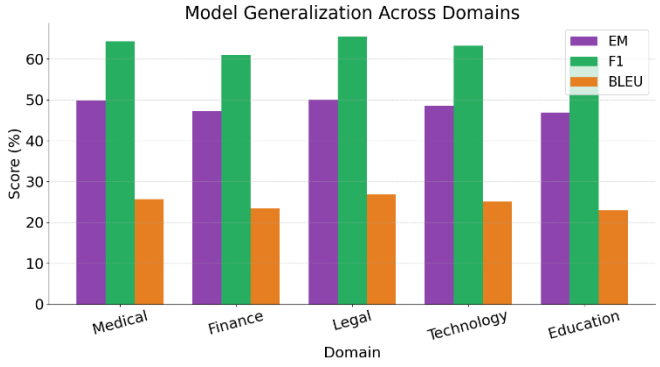

Figure 4. Verification experiment of model generalization ability in cross-domain scenarios.

 doi:10.20944/preprints202507.0826.v1

Figure 4 shows the generation performance of the proposed RAG model, which integrates structured and unstructured knowledge, across different domains. The model achieves relatively better results in domains with high specialization, such as medicine, law, and technology. In particular, the F1 score reaches 65.5% in the legal domain. This indicates that the model can accurately capture complex entity relationships and domain-specific semantic cues, demonstrating strong cross-domain generalization ability. This suggests that the knowledge fusion mechanism effectively supports high-quality answer generation even when dealing with diverse domain terminologies and expression styles. These results further confirm the stable role of structured knowledge in representing complex professional content, which is crucial for enhancing model adaptability. While the overall BLEU scores remain modest, the model achieves over 25% in the legal and technology domains, indicating a notable level of robustness in linguistic structure and syntactic expression. These results suggest that incorporating structured knowledge offers substantial linguistic support, effectively addressing common issues such as logical inconsistency in generative models when applied to unfamiliar domains. Overall, the experimental findings demonstrate that the proposed fusion mechanism enhances the accuracy of knowledge-based generation and exhibits strong transferability and generalization across domains. The model's stable performance in cross-domain scenarios underscores its broad applicability and practical value, particularly in multi-context environments that demand high-precision question answering.

5. Conclusions

This study proposes a RAG-based generation method that integrates structured and unstructured knowledge. The goal is to improve the accuracy and generality of large language models in factual question-answering and knowledge-intensive tasks. By constructing a dual-channel knowledge retrieval mechanism and a unified knowledge fusion network, the model effectively leverages the strengths of both knowledge types. It enhances factual support and logical consistency while maintaining the fluency of language generation. Experimental results show that the proposed method outperforms existing public models across multiple standard metrics. It also demonstrates stronger robustness and generalization in low-resource and cross-domain settings.

Compared to traditional generation systems that rely solely on unstructured documents, this method introduces structured knowledge to improve the handling of complex entity relations, multi-hop reasoning, and fine-grained question matching. Structured knowledge provides clearer semantic boundaries and conceptual organization. It also fills gaps in the pre-trained knowledge of the model. This advantage is particularly evident in specialized domains. Through detailed experiments on different types and proportions of structured knowledge, this study further validates the controllability and scalability of the fusion strategy. It offers both theoretical and practical support for building modular and domain-adaptive generation systems.

The findings of this study have broad application value in knowledge-intensive scenarios such as medical consultation, legal question answering, financial analysis, and educational tutoring. In these areas, models require higher accuracy and more precise terminology than in general open-domain generation tasks. The proposed fusion method enhances the domain expertise of QA systems. It also lays a solid foundation for the development of high-reliability and high-consistency human-computer interaction systems. In addition, the method shows strong compatibility and can be seamlessly integrated with existing language models and retrieval frameworks. This contributes to the advancement of multimodal and multi-source intelligent systems.

6. Future Work

Looking forward, as large models evolve toward higher parameter scales and stronger generalization capabilities, several future directions deserve attention. These include optimizing the representation of structured knowledge, introducing dynamic knowledge update mechanisms, and exploring the role of multimodal structured information, such as charts and flow diagrams, in

 doi:10.20944/preprints202507.0826.v1

generation tasks. At the same time, improving computational efficiency while maintaining performance will be key to enabling large-scale deployment and real-world applications. Future research may focus on adaptive knowledge enhancement designs to expand the model's applicability in complex scenarios such as multilingual, cross-cultural, and real-time interactive environments.

References

1. P. Zhao, H. Zhang, Q. Yu, et al., "Retrieval-augmented generation for AI-generated content: a survey," arXiv preprint, arXiv:2402.19473, 2024.
2. X. Li, J. Jin, Y. Zhou, et al., "From matching to generation: a survey on generative information retrieval," ACM Transactions on Information Systems, vol. 43, no. 3, pp. 1–62, 2025.
3. H. Wang, Y. Liu, C. Zhu, et al., "Retrieval enhanced model for commonsense generation," arXiv preprint, arXiv:2105.11174, 2021.
4. Z. Shao, Y. Gong, Y. Shen, et al., "Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy," arXiv preprint, arXiv:2305.15294, 2023.
5. D. Gao, "Deep graph modeling for performance risk detection in structured data queries," Journal of Computer Technology and Software, vol. 4, no. 5, 2025.
6. X. Wang, "Time-aware and multi-source feature fusion for transformer-based medical text analysis," Transactions on Computational and Scientific Methods, vol. 4, no. 7, 2024.
7. Z. Jiang, X. Ma, and W. Chen, "LongRAG: enhancing retrieval-augmented generation with long-context LLMs," arXiv preprint, arXiv:2406.15319, 2024.
8. L. Zhu, F. Guo, G. Cai, and Y. Ma, "Structured preference modeling for reinforcement learning-based fine-tuning of large models," Journal of Computer Technology and Software, vol. 4, no. 4, 2025.
9. Y. Gao, Y. Xiong, X. Gao, et al., "Retrieval-augmented generation for large language models: a survey," arXiv preprint, arXiv:2312.10997, vol. 2, no. 1, 2023.
10. X. Zheng, Z. Weng, Y. Lyu, et al., "Retrieval augmented generation and understanding in vision: a survey and new outlook," arXiv preprint, arXiv:2503.18016, 2025.
11. Y. Huang and J. Huang, "A survey on retrieval augmented text generation for large language models," arXiv preprint, arXiv:2404.10981, 2024.
12. Y. Peng, "Context-aligned and evidence-based detection of hallucinations in large language model outputs," Transactions on Computational and Scientific Methods, vol. 5, no. 6, 2025.
13. Y. Xing, T. Yang, Y. Qi, M. Wei, Y. Cheng, and H. Xin, "Structured memory mechanisms for stable context representation in large language models," arXiv preprint, arXiv:2505.22921, 2025.
14. Y. Peng, "Structured knowledge integration and memory modeling in large language systems," Transactions on Computational and Scientific Methods, vol. 4, no. 10, 2024.
15. H. Zheng, L. Zhu, W. Cui, R. Pan, X. Yan, and Y. Xing, "Selective knowledge injection via adapter modules in large-scale language models," 2025.
16. H. Zhang, Y. Ma, S. Wang, G. Liu, and B. Zhu, "Graph-based spectral decomposition for parameter coordination in language model fine-tuning," arXiv preprint, arXiv:2504.19583, 2025.
17. H. Zheng, Y. Wang, R. Pan, G. Liu, B. Zhu, and H. Zhang, "Structured gradient guidance for few-shot adaptation in large language models," arXiv preprint, arXiv:2506.00726, 2025.
18. F. Guo, L. Zhu, Y. Wang, and G. Cai, "Perception-guided structural framework for large language model design", 2025.
19. W. Zhang, Z. Xu, Y. Tian, Y. Wu, M. Wang, and X. Meng, "Unified instruction encoding and gradient coordination for multi-task language models," 2025.
20. Y. Deng, "Transfer methods for large language models in low-resource text generation tasks," Journal of Computer Science and Software Applications, vol. 4, no. 6, 2024.
21. T. Tang, "A meta-learning framework for cross-service elastic scaling in cloud environments", 2024.
22. Y. Ma, G. Cai, F. Guo, Z. Fang, and X. Wang, "Knowledge-informed policy structuring for multi-agent collaboration using language models," Journal of Computer Science and Software Applications, vol. 5, no. 5, 2025.

 doi:10.20944/preprints202507.0826.v1

23. Y. Xing, "Bootstrapped structural prompting for analogical reasoning in pretrained language models," Transactions on Computational and Scientific Methods, vol. 4, no. 11, 2024.
24. H. Xin and R. Pan, "Self-attention-based modeling of multi-source metrics for performance trend prediction in cloud systems," Journal of Computer Technology and Software, vol. 4, no. 4, 2025.
25. C. Liu, B. Wang, and Y. Li, "Dialog generation model based on variational Bayesian knowledge retrieval method," Neurocomputing, vol. 561, p. 126878, 2023.
26. S. Hofstätter, J. Chen, K. Raman, et al., "FiD-Light: efficient and effective retrieval-augmented text generation," Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1437–1447, 2023.
27. S. AboulEla, P. Zabihitari, N. Ibrahim, et al., "Exploring RAG solutions to reduce hallucinations in LLMs," Proceedings of the 2025 IEEE International Systems Conference (SysCon), pp. 1–8, 2025.
28. J. R. A. .Moniz, S. Krishnan, M. Ozyildirim, et al., "ReALM: reference resolution as language modeling," arXiv preprint, arXiv:2403.20329, 2024