Transactions on Computational and Scientific Methods | Vol. 4, No. 10, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Structured Knowledge Integration and Memory Modeling in Large Language Systems

Yuting Peng
New York University, New York, USA
yp2212@nyu.edu

Abstract: This study addresses the limitations of current large language models in long-term dependency retention and structured knowledge modeling. A fine-tuning algorithm is proposed, integrating memory networks and perception graph mechanisms, to improve overall performance in multi-hop reasoning and complex semantic understanding tasks. The method introduces a dynamic, readable and writable external memory module. It effectively stores and retrieves historical semantic information, alleviating the forgetting issue in long-text processing. At the same time, a perception graph is constructed to represent multi-dimensional relations among entities. A graph neural network is used to encode the graph structure, enabling deep integration between structured knowledge and the semantic space of the language model. Experiments are conducted on the HotpotQA dataset, covering samples with varying reasoning difficulties. Results show that the enhanced model outperforms the baseline in F1 score, semantic consistency, and reasoning stability. This confirms the effectiveness of the proposed fusion mechanism in complex language tasks. Further comparative experiments examine the impact of different graph neural network architectures on graph encoding performance. The results highlight the critical role of model architecture in the fusion mechanism. This study provides a technical approach to enhance knowledge retention and multi-level semantic modeling in large language models.

Keywords: Large language models, memory networks, perceptual maps, fine-tuning

1. Introduction

With the continuous development of artificial intelligence, Large Language Models (LLMs) have become central tools in natural language processing. They demonstrate strong capabilities in language understanding and generation across tasks such as dialogue generation, information extraction, and automatic question answering. In recent years, pre-trained language models such as GPT, BERT, and T5 have gained mainstream adoption. These models acquire deep understanding of linguistic structure and semantics through large-scale corpus pre-training. However, their performance remains limited by the scope of training data and their inherent generality. In domain-specific or complex tasks, they often face challenges such as insufficient semantic depth, limited reasoning ability, and degraded context retention. Therefore, enhancing LLMs' memory of long-term semantic dependencies and introducing structured knowledge to improve reasoning and perception has become a key research direction[1].

Memory Networks, as a mechanism for explicitly modeling long-term dependencies, provide an extended pathway for integrating external memory modules into LLMs. By introducing dynamic memory units that support read and write operations, the model can flexibly access historical information during multi-step reasoning or complex instruction scenarios. This effectively alleviates the "forgetting" problem when

processing long texts or multi-turn dialogues. Moreover, the interpretability of memory networks offers theoretical support for tracing the logic of model outputs. This is especially valuable in decision-sensitive and knowledge-intensive tasks. However, traditional memory mechanisms typically store information in fragmented and unstructured forms. They lack abstract modeling of inter-knowledge relationships, which restricts their reasoning capacity in complex semantic scenarios[2].

Meanwhile, the Perception Graph, as a multimodal knowledge graph integrating vision, language, and structured information, offers a way to organize and express cognitive data effectively. It models multi-dimensional relationships between conceptual entities through a graph structure. This enables language models not only to acquire specific information but also to understand higher-order associations. As a result, they gain human-like semantic association and contextual awareness. Incorporating the perception graph into the fine-tuning process of language models can enhance structural cognition without compromising generative flexibility. It significantly improves performance in tasks such as knowledge reasoning, event inference, and semantic aggregation[3]. Additionally, perception graphs are inherently compatible with external knowledge bases, task-oriented knowledge, and situational graphs, providing a more complete foundation for cognition.

Against this background, fine-tuning algorithms for LLMs that integrate memory networks and perception graphs have emerged. These aim to build enhanced language systems with both memory retention and knowledge perception capabilities. This integration addresses the limitations of traditional models in knowledge retention and semantic coherence. It also enables deep semantic modeling and dynamic reasoning based on specific task requirements while preserving general architecture. Particularly in complex multi-turn interactions, knowledge-intensive question answering, or context-dependent generation, the integrated mechanism significantly improves overall performance, usability, and scalability. By combining structured graphs with dynamic memory, it becomes possible to model human-like cognitive pathways and empower intelligent systems with advanced semantic understanding[4,5].

Therefore, this study aims to explore fine-tuning methods for LLMs based on the integration of memory networks and perception graphs. Inspired by human cognitive structures, we propose a multi-level, interactive, and structured language understanding framework. By constructing an efficient graph-memory fusion architecture, this research addresses bottlenecks in current large models for specific tasks. It also lays a foundation for building the next generation of general-purpose AI systems with high reliability and interpretability. Theoretically, it promotes the integration of cognitive computing and natural language processing. Practically, it offers feasible technical paths for intelligent question answering systems, decision support systems, and knowledge-based search engines, demonstrating strong theoretical significance and application potential.

2. Method

This study proposes a multi-level fine-tuning framework that integrates memory networks and perception graphs to enhance the knowledge expression, long-term memory retention, and structured reasoning capabilities of large language models in complex semantic tasks. The overall method is based on the basic large language model and introduces two external enhancement mechanisms: one is a dynamic memory network built on readable and writable memory units, which is used to achieve persistent modeling of historical semantic content; the other is a perception graph built on a multimodal semantic graph, which is used to guide the language model to establish a structured mapping between semantic relations. The framework achieves collaborative training of three sub-modules through a joint optimization strategy, and realizes efficient fusion and expression transfer of semantic information through perceptual alignment and attention scheduling mechanisms. The model architecture is shown in Figure 1.

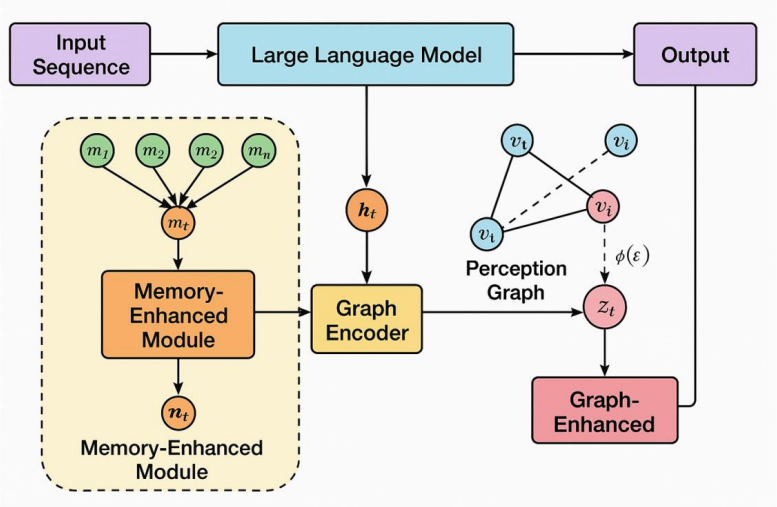


Figure 1. Overall model architecture

Assume that the basic language model is a pre-trained model $f_\theta(x)$, where x is the input sequence and θ is the model parameter. In the fine-tuning stage, we introduce the memory enhancement module $M = \{m_1, m_2, ..., m_n\}$, where each memory unit $m_i \in R^d$ represents the vector representation of a historical semantic fragment. After the semantic encoding of the current input x_t is $h_t = f_\theta(x_t)$, the matching degree between it and the memory unit is calculated through the attention mechanism:

$$a_i = \frac{\exp(h_t^T m_i)}{\sum_{j=1}^{n} \exp(h_i^T m_j)}$$

Get the weighted memory vector:

$$m_t = \sum_{i=1}^{n} a_i m_i$$

Finally, the memory vector is concatenated with the current semantic vector as the enhanced input $h_t^{'} = [h_t; m_t]$ and sent to the downstream task prediction module to realize the fusion utilization of historical information.

On the other hand, the perceptual graph structure is defined as a directed graph $G = (V, E)$, where the node set V represents the conceptual entity and the edge set $E \subseteq V \times V$ represents the semantic relationship. Each node V_i is embedded as a vector $e_i \in R^d$, and the structure is propagated and aggregated through the graph neural network (GNN). Specifically, the graph attention mechanism is used to calculate the information weight between adjacent nodes:

$$\beta_{i,j} = \frac{\exp(LeakyRELU(a^T[We_i \| We_j]))}{\sum_{k \in N(i)} \exp(LeakyRELU(a^T[We_i \| We_k]))}$$

Aggregation results in node enhancement representation:

$$e_i^{'} = \sum_{j \in N(i)} \beta_{ij} We_j$$

The graph semantic vector is then projected into the language model semantic space through the mapping function $\phi(e_i^{'})$ and input into the fusion module together with the memory-enhanced representation to achieve multi-source alignment of knowledge and semantics.

In the final fusion stage, we introduce a dynamic gating mechanism to regulate the fusion weight of memory representation and graph representation. Given the current language model representation h_t, memory representation m_t, and graph projection representation $\phi(e_i^{'})$, we define the fusion representation as:

$$z_t = \gamma \cdot h_t + \lambda \cdot m_t + (1-\gamma-\lambda) \cdot \phi(e_i^{'})$$

Where λ、γ is a learnable parameter that is dynamically updated through back-propagation to achieve the optimal semantic combination. The final representation is fed into the downstream decoder to generate output or classification prediction.

The innovation of this method is to break the limitation of traditional language models that only rely on internal representation modeling, and guide the model to build a human-like cognitive path through external memory mechanism and structural graph information, which not only realizes the dynamic preservation of semantics, but also improves the model's structured understanding ability. In multiple language understanding and generation tasks, this mechanism shows better generalization performance and interpretability, and provides a new paradigm for controllable fine-tuning and semantic reasoning of large language models.

3. Experiment

3.1 Datasets

This study adopts HotpotQA as the primary dataset for model fine-tuning and evaluation. HotpotQA is a large-scale English question answering dataset designed for complex reasoning tasks. It was jointly developed by Carnegie Mellon University and Stanford University. The dataset contains over 110,000 questions requiring multi-document reasoning. Each sample includes a natural language question, several background documents, and a specific answer. Some samples also provide the sentence-level supporting facts. These features make HotpotQA suitable for testing both language comprehension and challenging multi-hop reasoning and semantic retention, aligning well with this study's focus on structured semantic modeling and long-term memory.

Unlike traditional single-turn QA datasets, HotpotQA emphasizes reasoning chains and cross-paragraph knowledge integration. Questions often require reasoning across multiple entity relations or concept hierarchies. To answer correctly, models must demonstrate strong capabilities in information filtering, relationship modeling, and logical composition. These characteristics make HotpotQA an ideal platform for evaluating multimodal fusion and graph-enhanced mechanisms. It is especially suitable for testing the effect of perception graphs on structural hierarchy awareness in language modeling. In addition, the dataset's "bridge entity" paths offer natural support for graph construction. They also enable memory networks to store and reuse intermediate reasoning results.

During data processing, we applied standardized preprocessing to the HotpotQA dataset. This includes text normalization, entity recognition, knowledge graph embedding alignment, and memory node construction. All inputs were encoded into formats compatible with the language model. Graph and memory modules were incorporated through additional channels to form a complete fused input. As a result, this dataset supported both the fine-tuning of various modules and served as the main evaluation benchmark. It allowed comprehensive testing of the model's performance in complex question answering and its capacity for knowledge retention.

3.2 Experimental Results

During the experiment, this paper first compares the impact of different graph neural network structures on graph encoding performance. The experimental results are shown in Table 1.

Table 1: Comparison of the impact of different graph neural network structures on graph encoding performance

Model	Node representation accuracy (Accuracy)	Cosine Similarity	Multi-hop question answering F1 value
GCN	82.4	0.743	71.2
GAT[6]	85.7	0.786	74.5
GraphSAGE[7]	84.1	0.765	73.1
RGAT[8]	87.3	0.802	76.4
HGT	88.9	0.816	78.0

Experimental results reveal distinct variations in performance among different graph neural network architectures when addressing the graph encoding task. In particular, the Heterogeneous Graph Transformer (HGT) emerged as the top-performing model across the board, surpassing other methods on every evaluation metric. Notably, it attained an 88.9% accuracy in node representation, alongside a graph consistency score of 0.816, which underscores its robust capacity to encode and preserve semantic coherence. Moreover, in the realm of multi-hop question answering, HGT's F1 score climbed to 78.0, further demonstrating its ability to capture and leverage complex semantic relationships within heterogeneous graph structures. These findings strongly suggest that more advanced GNN frameworks, particularly those that account for multiple types of relations and incorporate intricate modeling strategies, can significantly elevate performance in downstream language understanding tasks.

By contrast, established approaches such as GCN and GraphSAGE provide benefits in terms of computational speed and resource usage, yet they fall short in overall performance. For example, GCN manages only 82.4% accuracy in node representation, a figure that is notably below both HGT and RGAT. This deficit can be traced to GCN's simplistic neighbor aggregation strategy, which diminishes its capacity to capture more intricate semantic relationships. While GraphSAGE ameliorates some of these shortcomings by employing a neighbor-sampling approach during aggregation, it still lags behind attention-based models in graph consistency and multi-hop QA performance. Meanwhile, GAT and RGAT significantly strengthen graph representation by utilizing attention mechanisms. In particular, RGAT stands out for its aptitude in modeling directed edges and multi-relational graphs, thereby validating its sensitivity to structural nuances in the data.

Overall, when we employ more sophisticated and complex graph neural network architectures, the ability to capture richer semantic information and provide stronger support for language models increases correspondingly. Although advanced models such as HGT require substantially higher computational resources, their notable performance improvements make them particularly valuable for tasks that demand robust reasoning and multi-hop information integration. The findings from this experiment underscore the critical role of selecting an appropriate GNN structure to maximize the effectiveness of graph-enhanced language models. In addition, these results offer constructive guidance on how future work can further integrate and optimize different modules and architectural components for even better performance.

Next, this paper presents a detailed examination of how multi-hop reasoning accuracy is enhanced after introducing perceptual graphs, and demonstrates these improvements through empirical evidence. The

resulting data, illustrated in Figure 2, highlights the positive impact of perceptual graphs in boosting multi-hop reasoning performance.

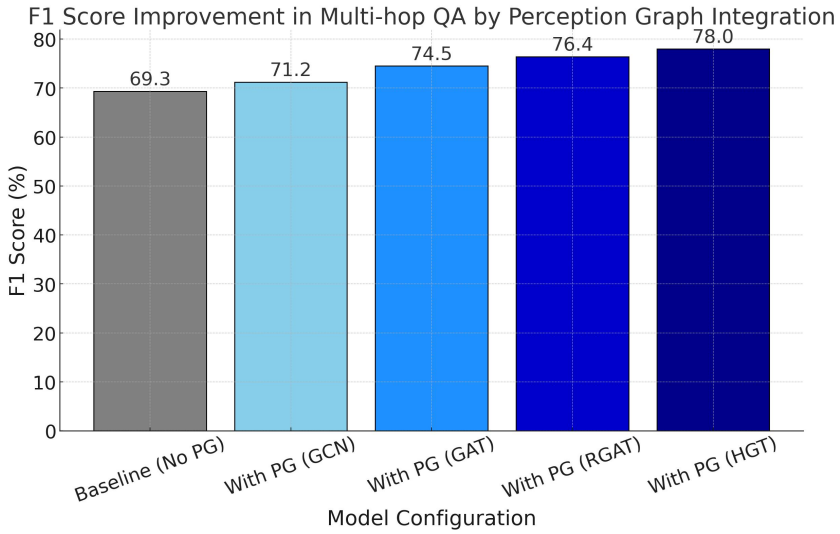


Figure 2. F1 Score Improvement in Multi-hop QA by Perception Graph Integration

As shown in the results of Figure 2, introducing the perception graph leads to a significant improvement in F1 scores on the multi-hop reasoning task. The baseline model, without any graph structure, achieved an F1 score of 69.3%. When supported by different graph neural network architectures, the performance improved across the board. Among them, HGT reached the highest score of 78.0%. This indicates that the perception graph, as an external structured semantic enhancement, can effectively guide large language models to build deeper semantic associations and reasoning paths.

When comparing different graph neural networks, it is clear that improved graph modeling capabilities lead to higher reasoning accuracy. For example, GCN relies on simple adjacency aggregation and achieved a limited improvement, reaching only 71.2%. In contrast, GAT and RGAT introduced attention mechanisms, allowing the model to better capture important semantic connections between nodes. Their F1 scores reached 74.5% and 76.4%, respectively. HGT further incorporated heterogeneous graph modeling, significantly enhancing the model's ability to process diverse relations. This led to the best overall performance.

This experiment clearly demonstrates that the incorporation of graph-based information can effectively mitigate the inherent limitations present in the native semantic representations of traditional language models. In particular, such graph-enhanced approaches significantly boost the models' generalization capabilities and reasoning performance in tasks such as multi-hop question answering, where a deep understanding of semantic structure is critical for success. Furthermore, the observed performance improvements are especially pronounced when employing more advanced and complex graph neural network architectures, which are capable of modeling intricate relationships and higher-order connections. These results provide robust empirical evidence in support of further research aimed at refining fusion mechanisms and optimizing the structural design of graph-enhanced language models.

Finally, this paper presents the results of a robustness evaluation, in which the model was tested on HotpotQA samples spanning various levels of difficulty. This experiment was conducted to assess the model’s stability and adaptability under more challenging conditions and to determine its effectiveness across a broader spectrum of question complexities. The detailed outcomes of this robustness test are visually illustrated in Figure 3, offering further insight into the model’s resilience and reliability when handling semantically demanding multi-hop reasoning tasks.

Figure 3 presents the robustness evaluation results of the model on the HotpotQA dataset under different difficulty levels. It is evident that the enhanced model consistently outperforms the baseline across all levels,

demonstrating stronger stability and adaptability. On Easy samples, the F1 score difference between the two models is 3.6 percentage points. On Hard samples, the gap widens to 5.9, indicating that the enhancement mechanism offers more significant advantages when dealing with complex semantics and cross-document reasoning. Notably, for Bridge-type questions, the enhanced model maintains high performance at 78.5%, while the baseline drops sharply to 73.8%. This highlights the role of structured semantic fusion in improving cross-semantic bridging ability.

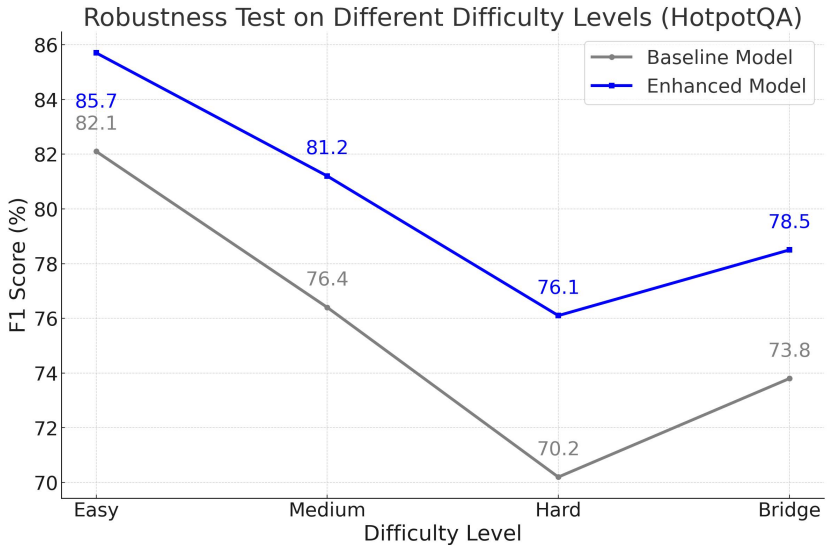


Figure 3. Robustness Test on Different Difficulty Levels

Overall, as task difficulty increases, the baseline model shows a more pronounced decline in performance. In contrast, the enhanced model, equipped with memory networks and perception graphs, maintains stable results across various complex reasoning scenarios. These findings confirm that the proposed model exhibits stronger robustness and generalization when facing questions involving semantic jumps and longer logical chains. This provides solid support for its practical application in handling diverse language inputs in real-world environments.

4. Conclusion

This paper proposes a fine-tuning algorithm for large language models that integrates memory networks and perception graph mechanisms. The goal is to enhance the model's ability to capture long-term dependencies and perceive structured knowledge in complex semantic tasks. By introducing an external memory module and a graph-based encoder, the model can dynamically retain historical semantic content and perform multi-level modeling of entity relations. As a result, the language model demonstrates stronger reasoning ability and semantic coherence in tasks such as multi-hop question answering. Experimental results show that on the HotpotQA dataset, the integrated model outperforms traditional large language models across different difficulty levels, achieving better accuracy and robustness. This study presents a structural fusion design that organically combines memory and graph-based enhancements to support human-like cognitive pathways in language modeling. Specifically, an attention-driven dynamic memory scheduling mechanism and a perception-aligned graph aggregation network were constructed. These components jointly improve the model's capacity to understand and retain semantic structures, entity relationships, and reasoning chains. Extensive comparative experiments validate the stability and transferability of the proposed method under various model architectures and graph configurations. The results lay a solid technical foundation for future research on structure-enhanced language models.

In addition, the experiments highlight the critical role of graph neural network structures in semantic graph modeling. Different graph designs significantly affect the performance of language models. The improved

results on complex samples indicate the value of combining semantic structure modeling with memory mechanisms in multi-hop reasoning tasks. This finding offers a promising direction for developing more efficient semantic graph representations. It also suggests that future fine-tuning strategies should consider dynamic alignment among model architecture, knowledge type, and task semantics. Future work can be extended in two directions. First, it is worth exploring how to embed multimodal perception information (such as vision and audio) into the graph-enhanced module in a unified structural form, enabling cross-modal language modeling. Second, to address large-scale open-domain QA and reasoning generation tasks, more efficient and generalized language model architectures should be developed. These models should incorporate mechanisms like federated learning and incremental learning to support continuous knowledge adaptation and update. This would provide key algorithmic support for general-purpose intelligent systems and further advance the development of human-like language understanding systems.

References

[1] Huang, Shaohan, et al. "Language is not all you need: Aligning perception with language models." Advances in Neural Information Processing Systems 36 (2023): 72096-72109.

[2] Perin, Wagner de A., et al. "From text to maps: Automated concept map generation using fine-tuned large language model." Simpósio Brasileiro de Informática na Educação (SBIE). SBC, 2023.

[3] Patel, Roma, and Ellie Pavlick. "Mapping language models to grounded conceptual spaces." International conference on learning representations. 2022.

[4] Chang, Qinglong, Kwok-Wai Hung, and Jianmin Jiang. "Fine tuning of deep contexts toward improved perceptual quality of in-paintings." IEEE Transactions on Cybernetics 52.6 (2021): 4850-4854.

[5] Kang, Sing Bing, and Katsushi Ikeuchi. "Toward automatic robot instruction from perception-mapping human grasps to manipulator grasps." IEEE transactions on robotics and automation 13.1 (1997): 81-95.

[6] Yang, L., Huang, B., Li, Q., Tsai, Y.-Y., Lee, W. W., Song, C., & Pan, J. (2023). TacGNN: Learning Tactile-based In-hand Manipulation with a Blind Robot. arXiv preprint arXiv:2304.00736.

[7] Zhu, Wenhui, et al. "An Expert Data Generation Method for Multi-Agent Cooperative Planning Method." 2023 28th International Conference on Automation and Computing (ICAC). IEEE, 2023.

[8] Wang, X., Ji, H., Shi, C., Wang, B., Cui, P., Yu, P. S., & Ye, Y. (2019). Heterogeneous Graph Attention Network. Proceedings of the 2019 World Wide Web Conference (WWW '19), 2022–2032.