Graph-Guided Prompting for Zero-Shot Multi-Hop Question Generation: Gains without Fine-Tuning, Limits without Adaptation

Samin Jamshidi [1] **Morteza Mahdiani** [2 3]

Abstract

We propose a zero-shot framework for multi-hop question generation that couples a lightweight Graph Attention Network (GAT) with pretrained large language models. The GAT is trained to identify the entities most indicative of the reasoning chain within a passage–answer pair and to propagate relational information across the resulting entity graph. These predicted entities are then woven back into the passage, forming an entity-enriched prompt that is fed directly to existing language models, specifically LLAMA-2-7B and DEEPSEEK-CODER-6.7B, though the approach is extensible to newer LLMs with longer context windows. This decoupled design lets a single reasoning module enhance diverse language models at negligible computational cost. While we test on two open-source models, the modular nature of our framework allows for application to larger and newer models without architectural changes. Preliminary results on HotpotQA show that the GAT-augmented prompts yield consistent improvements in answer containment, syntactic diversity, and automatic metrics such as BLEU and ROUGE-L over plain zero-shot prompting and joint-training baselines. At the same time, performance still trails that of fully fine-tuned task-specific systems, suggesting that structured entity reasoning is complementary rather than a complete substitute to end-to-end adaptation.

1. Introduction

Multi-hop question generation (MHQG) aims to produce questions that require reasoning over multiple, dispersed evidence fragments within a document or across several documents. By demanding bridge inferences, entity linking, and cross-sentence synthesis, MHQG probes a deeper level of text understanding than single-hop generation. This capability underpins advanced reading-comprehension tools, curriculum design in intelligent tutoring systems, and open-domain question-answering pipelines.

Although recent large language models (LLMs) such as LLAMA-2-CHAT-7B (Touvron et al., 2023)and DEEPSEEK-CODER-INSTRUCT-6.7B (Bi et al., 2024) excel at surface-level generation, they often falter when asked to discover and integrate supporting facts for complex questions without external guidance (Lin et al., 2024). Benchmarks like HotpotQA (Yang et al., 2018) expose these limitations, showing that purely sequence-based models struggle with multi-hop phenomena such as bridge-entity reasoning.

Graph-based approaches have therefore been explored to make inter-entity relations explicit. Fei et al. (2022) inject entity constraints to control hop complexity, while Xia et al. (2023) employ multi-level content planning via intermediate answer summaries. Jamshidi & Chali (2025) jointly co-trains a graph neural network with a seq2seq generator, but the resulting entanglement of reasoning and language modeling adds computational overhead and reduces modularity.

This paper asks: *(i) Can a reasoning module trained in isolation supply sufficient structure to improve MHQG when paired with a powerful, inference-only LLM? (ii) Does such decoupling retain generation fluency while enhancing factual grounding and multi-hop coherence?*

We answer these questions through a modular framework that first trains a Graph Attention Network (GAT) to identify the entities and relations most critical for a given passage–answer pair, then weaves these predictions into an *entity-enriched prompt*. The prompt is delivered, without further parameter updates, to an off-the-shelf LLM. This separation keeps the reasoning component lightweight and domain-adaptive, while preserving the generator's pretrained fluency.

Experiments show that adding graph-extracted entities consistently improves inference-only LLMs over plain prompt-

[1]Department of Mathematics and Computer Science, University of Lethbridge, Canada [2]Department of Computer Science and Operations Research, University of Montreal, Canada [3]Mila – Quebec AI Institute, Canada. Correspondence to: Samin Jamshidi <jamshidisamin73@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

ing baselines, although the overall performance remains below that of end-to-end fine-tuned models. Even so, our compact graph-informed module provides a lightweight, plug-and-play avenue for scaling reasoning-aware question generation without additional language-model training.

Throughout this paper, the terms *Llama* and *DeepSeek* specifically denote the models LLAMA-2-CHAT-7Band DEEPSEEK-CODER-INSTRUCT-6.7B, respectively. We treat these and any analogous variants as *inference-only pretrained LLMs*, meaning their parameters remain fixed for the entirety of our pipeline.

2. Related Work

Multi-hop question generation extends standard QG by requiring the integration of multiple evidence fragments to form coherent, reasoning-intensive questions. Datasets such as HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022) have been developed to benchmark this capability, emphasizing multi-sentence reasoning. Early approaches used sequence-to-sequence architectures augmented with retrieval or sentence selection modules, but often failed to guarantee multi-hop complexity. Models like MHQG-GNN (Song et al., 2020) introduced multi-hop relational graphs over the context to guide generation, while GRAFT-Net (Sun et al., 2019) enabled reasoning over hybrid text-graph representations.

To improve the integration of structured reasoning, graph-based approaches have been widely adopted. GNET-QG (Jamshidi & Chali, 2025) jointly trains a Graph Neural Network and decoder to propagate information between related entities. HeterGS (Zhang et al., 2021) further models heterogeneous relations among different node types, capturing richer reasoning signals. CQG (Fei et al., 2022) enforces the presence of bridge entities in the generated question to guarantee multi-hop depth, while (Xia et al., 2023) uses a two-stage process to plan and generate questions based on a full-answer summary composed of supporting facts.

More recent work explores how large language models (LLMs) can perform MHQG with minimal supervision. TASE-CoT (Lin et al., 2024) prompts LLMs using type-aware semantics and chain-of-thought reasoning for question generation in few-shot settings. Our method complements this trend by proposing a modular design: we train a Graph Attention Network (GAT) to extract entities and reason over them, then pass its output to a inference-only pretrained LLM (Llama or DeepSeek). This avoids the complexity of joint fine-tuning while retaining both reasoning ability and linguistic fluency.

Recent approaches such as TASE-CoT focus on enriching prompts with type-aware semantics and chain-of-thought reasoning in few-shot settings. While effective, these require prompt engineering at the task level. In contrast, our method separates structural reasoning via GAT, making it more adaptable across tasks and model variants.

3. Methodology

3.1. Graph Construction and Reasoning

We first construct an entity graph from the input document, where nodes correspond to named entities and edges represent semantic or structural relationships. Specifically, we connect entities based on three criteria: sentence-level co-occurrence, paragraph title association, and cross-paragraph coreference. Each node is initialized with a contextualized embedding derived from a pre-trained BART encoder (Lewis et al., 2019), enabling it to capture rich, bidirectional contextual information.

The graph is processed using a multi-head Graph Attention Network with 4 heads and a hidden size of 128. These were chosen based on preliminary performance and runtime trade-offs. BART embeddings were selected for their contextual richness in entity disambiguation. The GAT learns to attend over neighboring entities to refine node representations based on their relevance to the target question. We train the GAT using a binary classification objective, labeling as positive the entities that appear in both the supporting facts and the reference question span. This encourages the model to identify entities that act as reasoning bridges and are essential for generating complex questions. Figure 2 illustrates a monotonic decline in training loss and a consistent increase in validation F_1, which peaks at 0.288 in epoch 8.

3.2. Generating Enriched Context

After identifying the most relevant entities, we extract their updated GAT embeddings and convert them into their textual representations. These entities, denoted as E_{sub}, are concatenated with the original context C and the provided answer A to form the enriched input:

$$C_{\text{enriched}} = [C; A; E_{\text{sub}}] \tag{1}$$

This enriched context provides the language model with both global context and a focused reasoning signal, enhancing its ability to generate coherent and well-grounded multi-hop questions.

3.3. Prompting Inference-Only LLMs with Entity-Enriched Context

The enriched context C_{enriched} is fed as a text prompt to an inference-only pretrained LLM, decoder-only transformer model. We experiment with two state-of-the-art LLMs: Llama and DeepSeek, accessed via the Hugging Face Transformers library. The models generate questions autoregres-

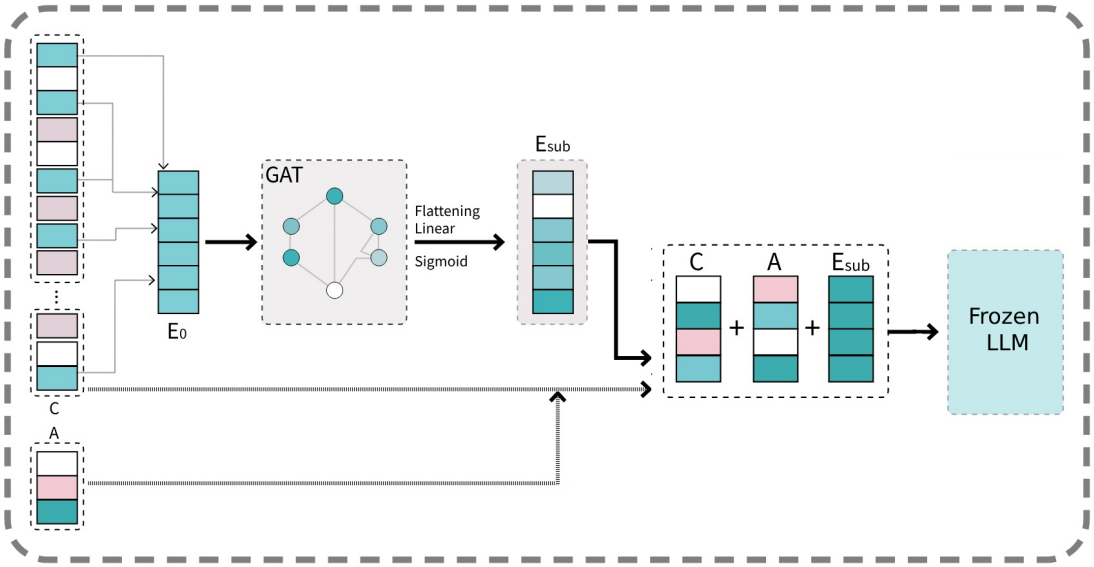

Figure 1. Overview of our proposed architecture. C represents the original context, and A denotes the given answer. Named entities are extracted from the context and passed through a Graph Attention Network (GAT), which performs relational reasoning over the entity graph. The output node embeddings are flattened, linearly transformed, and passed through a sigmoid layer to obtain importance scores. Entities with scores above a threshold are selected to form E_{sub}, the subset of salient reasoning entities. The enriched input context, composed of $[C; A; E_{\text{sub}}]$, is then passed to a inference-only pretrained LLM (e.g., Llama or DeepSeek) to generate the final multi-hop question.

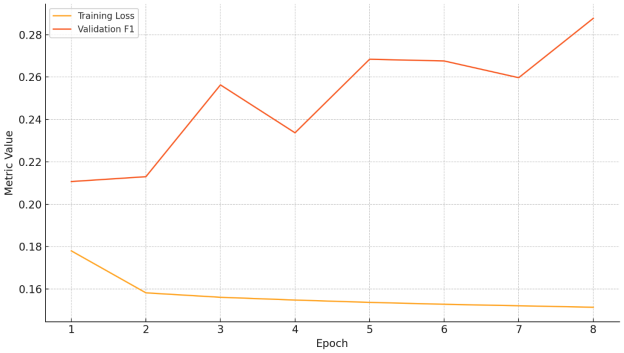

Figure 2. Training trajectory of the entity-selection Graph Attention Network.

sively using greedy decoding. No fine-tuning is applied; all parameters of the LLM remain frozen during inference.

This separation of responsibilities allows the reasoning module to evolve independently from the language model. Compared to GNET-QG (Jamshidi & Chali, 2025), our framework avoids the complexity of joint optimization and facilitates plug-and-play compatibility with newer LLMs. This modularity also significantly reduces computational overhead and improves generalizability across domains.

4. Results

To evaluate the effectiveness of our proposed GAT-enhanced question generation framework, we conducted automatic evaluations on the HotpotQA test set using standard metrics widely adopted in the question generation literature. Specifically, we report BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Lavie & Agarwal, 2007), which assess various aspects of fluency, precision, recall, and semantic alignment. These metrics enable us to compare the output of our model with strong baselines and prior multihop QG systems.

Table 1 shows that both *GAT + Llama* and *GAT + DeepSeek* outperform their respective base models across all automatic metrics on the HotpotQA dataset.

BLEU Scores: We observe consistent improvements from BLEU-1 to BLEU-4, indicating better lexical alignment with the reference questions, particularly for *GAT + DeepSeek*.

ROUGE-L: Gains of +1.55 (Llama) and +1.93 (DeepSeek) over baselines suggest better structural overlap and contextual recall in the generated questions.

METEOR: The highest relative gains (+1.5 for both) reflect improved semantic fluency and alignment.

Qualitative inspection reveals that GAT-selected entities help the LLM focus on key reasoning chains, resulting in

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
GAT + Llama	10.77	4.66	2.55	1.68	14.17	16.53
Llama	10.05	4.19	2.38	1.61	12.62	15.02
GAT + DeepSeek	16.98	8.36	4.75	3.11	21.87	26.59
DeepSeek	16.24	7.95	4.61	3.05	19.94	25.06

Table 1. Automatic evaluation on HotpotQA. Metrics (BLEU, ROUGE-L, METEOR) are reported for the base generators (LLAMA and DEEPSEEK) and for their graph-augmented counterparts (*GAT + Llama* and *GAT + DeepSeek*).

more coherent and specific multi-hop questions. While our models do not outperform end-to-end fine-tuned systems, they demonstrate that structured reasoning can be added effectively without updating any LLM parameters and at minimal computational cost.

5. Conclusion

We introduced a modular, graph-guided framework that supplies explicit entity-relation signals to inference-only pretrained LLMs for multi-hop question generation. The approach yields consistent gains over text-only prompting yet does not exceed the performance of bespoke, fully fine-tuned generators. This gap is unsurprising: our models remain fixed during inference, and the quality of generation is therefore bounded by (i) the expressive capacity of the underlying LLM and (ii) the accuracy of the upstream entity extractor. Nevertheless, the results demonstrate that structured reasoning can be injected at negligible training cost and with full model reuse, providing a strong foundation for future work that couples the same graph module with larger back-ends, optimized prompt templates, or a lightweight stage of task-specific adaptation. These directions hold promise for closing, and potentially reversing, the margin between parameter-locked and fine-tuned solutions in multi-hop question generation. We see this work as a step toward low-resource MHQG systems that can seamlessly plug into evolving LLM backbones with minimal intervention.

6. Limitations and future directions.

This study concentrates on parameter–locked, openly available LLMs of approximately seven billion parameters (LLAMA-2-CHAT-7B and DEEPSEEK-CODER-INSTRUCT-6.7B), reflecting the resource profile of many academic research groups. While these models already provide a strong test bed, the proposed framework is not limited to them. Future investigations can broaden the evaluation by (i) enriching the entity–enhanced prompt with a small set of few-shot exemplars, (ii) refining prompt templates through targeted, reasoning-oriented optimization, and (iii) applying the method to a wider range of model families and parameter scales, including larger or commercially licensed systems. Such extensions will help establish the generality and scalability of graph-guided question generation across diverse language-modeling regimes.

Our current implementation uses a fixed enriched prompt format where the context (C), the answer (A), and a subset of salient entities (Esub) are concatenated linearly as [C; A; Esub]. While this format consistently improves performance, we do not explore alternative insertion strategies or entity placements within the prompt. Future work could investigate learned or dynamically structured prompt templates tailored to different LLM architectures or reasoning styles, potentially improving alignment between injected structure and model behavior.

References

Bi, X. et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Fei, Z., Zhang, Q., Gui, T., Liang, D., Wang, S., Wu, W., and Huang, X. CQG: A simple and effective controlled generation framework for multi-hop question generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6896–6906, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.475. URL https://aclanthology.org/2022.acl-long.475.

Jamshidi, S. and Chali, Y. GNET-QG: Graph network for multi-hop question generation. In Gesese, G. A., Sack, H., Paulheim, H., Merono-Penuela, A., and Chen, L. (eds.), *Proceedings of the Workshop on Generative AI and Knowledge Graphs (GenAIK)*, pp. 20–26, Abu Dhabi, UAE, jan 2025. International Committee on Computational Linguistics. URL https://aclanthology.org/2025.genaik-1.3.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lavie, A. and Agarwal, A. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 228–231, Prague,

Czech Republic, jun 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0734.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. URL https://arxiv.org/abs/1910.13461.

Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, jul 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Lin, Z., Chen, W., Song, Y., and Zhang, Y. Prompting few-shot multi-hop question generation via comprehending type-aware semantics. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3730–3740, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.237. URL https://aclanthology.org/2024.findings-naacl.237.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.18653/v1/P02-1040. URL https://aclanthology.org/P02-1040.

Song, L., Wang, Z., Wang, Y., Zhang, Y., and Gildea, D. Multi-hop question generation with graph neural networks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3192–3202, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.286. URL https://aclanthology.org/2020.coling-main.286.

Sun, H., Dhingra, B., Zaheer, M., Liu, Z., Cohen, W. W., and Salakhutdinov, R. Open domain question answering using early fusion of knowledge bases and text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4231–4240, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1431. URL https://aclanthology.org/D19-1431.

Touvron, H. et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Trivedi, H., Zhang, N., Feigenblat, G., Xiong, C., Frieder, O., and McCallum, A. MuSiQue: Multihop questions via single-hop question composition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3467–3480, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.243. URL https://aclanthology.org/2022.acl-long.243.

Xia, Z., Gou, Q., Yu, B., Yu, H., Huang, F., Li, Y., and Nguyen, C.-T. Improving question generation with multi-level content planning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 800–814, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.54. URL https://aclanthology.org/2023.findings-emnlp.54.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint*, 2018. URL https://arxiv.org/abs/1809.09600.

Zhang, Y., Wang, Z., Zhang, Y., Zhou, Z., Yu, M., and Gildea, D. Heterogeneous graph structure learning for multi-hop question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5892–5904, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.476. URL https://aclanthology.org/2021.emnlp-main.476.