SOCIAL AND MANAGEMENT SCIENCES RESEARCH JOURNAL

www.smsrj.com

p-ISSN: 2791-0415 **e-ISSN:** 2791-0423

Volume:4 Issue: 02(2026)

Verifier-Mediated Debate for Controllable Question Synthesis over Knowledge Graphs

Lisa Weiß

Institute of Computer Science, Faculty of Mathematics and Natural Sciences, University of Bonn, Bonn, North Rhine-Westphalia, Germany

Abstract:

The rapid expansion of knowledge graphs has necessitated advanced interfaces capable of translating structured data into natural language queries, an area known as question synthesis. Traditional approaches to question synthesis over knowledge graphs often struggle with semantic fidelity, hallucination, and a lack of structural controllability, particularly as the complexity of the underlying subgraphs increases. This paper introduces a novel framework, Verifier-Mediated Debate, designed to address these challenges by employing a multi-agent debate mechanism strictly regulated by a deterministic verifier. By casting question generation as a discursive process between opposing generative agents, the proposed system iteratively refines syntactic coherence and structural alignment. A neutral verifying agent, grounded explicitly in the source knowledge graph schema, objectively evaluates proposed questions at each debate round, terminating the discourse once optimal semantic fidelity is achieved. Extensive theoretical analysis and empirical evaluations demonstrate that this framework significantly enhances the controllability and accuracy of synthesized questions compared to single-agent generative paradigms. The integration of a deterministic verifier curtails the introduction of spurious information, thereby resolving a critical bottleneck in large language model applications for structured data. This research advances the paradigm of neuro-symbolic integration and establishes a robust foundation for building trustworthy, automated question generation systems tailored for complex knowledge repositories.

Keywords: *Knowledge Graphs, Question Synthesis, Multi-Agent Debate, Controllable Generation*

1. Introduction

The proliferation of structured data repositories, commonly formalized as knowledge graphs, has transformed the landscape of information retrieval, natural language processing, and automated reasoning [1]. Knowledge graphs encode factual information in the form of interconnected entities and relations, providing a rich, machine-readable substrate for complex querying operations. However, the semantic chasm between structured triplets and human-readable natural language remains a significant barrier to the widespread accessibility of these repositories [2]. Question synthesis over knowledge graphs, which aims to automatically generate natural language questions corresponding to specific subgraphs or logical forms, has emerged as a crucial task to bridge this gap [3]. This task not only facilitates the creation of large-scale training datasets for semantic parsing and question answering systems but also serves as a foundational component for conversational agents designed to interact with structured knowledge [4].

Despite substantial progress driven by the advent of deep learning architectures, contemporary question synthesis systems face persistent challenges regarding controllability and semantic fidelity. Early sequence-to-sequence models, while effective for simple single-hop facts, frequently falter when tasked with generating multi-hop questions requiring complex logical reasoning. The integration of large language models has partially mitigated these syntactic shortcomings; however, these models are notoriously prone to hallucination, wherein the generated text includes entities or constraints absent from the source subgraph [5]. Furthermore, purely generative approaches lack intrinsic mechanisms for fine-grained control over the structural complexity, target reasoning type, or lexical diversity of the synthesized questions [6]. This lack of controllability restricts the utility of synthetic data in scenarios demanding strict adherence to specific logical templates or domain constraints. To overcome these limitations, this study proposes a paradigm shift from monolithic generation to an interactive, multi-agent framework termed Verifier-Mediated Debate. The core premise of this approach relies on the instantiation of two distinct generative agents tasked with iteratively proposing and critiquing natural language questions derived from a given input subgraph [7]. This adversarial, discursive process is designed to expose syntactic flaws and semantic drift. However, unbounded debate among generative models often leads to divergence or the collective reinforcement of erroneous assumptions. Therefore, a critical innovation of the proposed architecture is the introduction of a deterministic verifying agent. This verifier grounds the debate strictly in the original knowledge graph schema, objectively evaluating each proposed question against the input constraints [8]. The verifier mediates the discourse, providing targeted feedback to the generative agents and terminating the debate when a predefined threshold of structural and semantic alignment is achieved. The primary objective of this paper is to rigorously define, formalize, and evaluate the Verifier-Mediated Debate framework for controllable question synthesis. This study systematically deconstructs the generation process into constituent debate rounds, mathematically formulating the reward mechanisms and update rules governing the generative agents. By enforcing strict verifier mediation, the framework effectively constrains the hypothesis space, significantly reducing the incidence of hallucination and ensuring high-fidelity outputs [9]. The methodology is designed to be agnostic to the specific architecture of the underlying large language models, offering a highly generalized solution applicable across various domains and knowledge graph schemas. In the subsequent sections, this paper outlines a comprehensive theoretical framework and empirical evaluation of the proposed system. Section 2 provides an exhaustive review of the relevant literature encompassing knowledge graphs, question synthesis, and multi-agent debate mechanisms in artificial intelligence [10]. Section 3 details the conceptual architecture of the Verifier-Mediated Debate framework,

establishing the mathematical formulations underpinning the multi-agent interactions. Section 4 elaborates on the specific methodologies employed for controllable question generation, including subgraph extraction and verification protocols. Section 5 describes the experimental design, dataset selection, and evaluation metrics utilized to validate the framework. Section 6 presents a detailed analysis of the quantitative and qualitative results, while Section 7 concludes the paper by summarizing the contributions and identifying promising directions for future research in the domain of neuro-symbolic text generation.

2. Background and Literature Review

2.1 Evolution of Question Synthesis over Knowledge Graphs

The task of question synthesis, also known as question generation, has evolved significantly over the past decade, mirroring broader advancements in natural language processing. Initially, question generation over structured data relied heavily on rule-based systems and deterministic templates. These early approaches required extensive manual labor from domain experts to craft grammatical rules and mapping functions that translated specific graph motifs into corresponding natural language templates [11]. While highly controllable and free from hallucination, template-based systems suffered from a severe lack of linguistic diversity and scalability, proving inadequate for handling the heterogeneous schema structures typical of modern, open-domain knowledge graphs [12].

The paradigm shifted dramatically with the introduction of neural sequence-to-sequence models, which framed question synthesis as a translation task from linearized graph structures to natural language sequences. Recurrent neural networks, augmented with attention mechanisms, enabled models to align specific tokens in the output sequence with distinct entities or relations in the input subgraph. This neural approach significantly improved the fluency and diversity of the generated questions. However, flattening a graph into a sequence inherently discards valuable topological information [13]. Subsequent research introduced graph neural networks to better encode the multi-relational structure of the input data before passing the representations to a text decoder [14]. Graph-to-sequence models demonstrated superior capability in synthesizing multi-hop questions, as they could explicitly model the structural dependencies between distant entities in the subgraph. Despite these architectural improvements, issues of semantic drift and logical inconsistency persisted. Neural models, trained via maximum likelihood estimation, often prioritized statistical correlations in the training data over strict adherence to the input facts [15]. This led to the generation of plausible-sounding but factually incorrect questions, a problem that became particularly pronounced when synthesizing complex queries involving aggregations, comparative reasoning, or temporal constraints [16]. Consequently, the field recognized a pressing need for mechanisms that could enforce semantic fidelity and structural alignment during the generation process.

2.2 Large Language Models and the Hallucination Bottleneck

The contemporary era of natural language processing is dominated by large language models based on the transformer architecture. These models, pre-trained on vast corpora of text, exhibit remarkable few-shot and zero-shot generation capabilities, capturing complex linguistic patterns and broad encyclopedic knowledge. When applied to question synthesis, large language models can generate highly fluent and contextually rich questions from minimal prompts [17]. Researchers have leveraged prompt engineering and in-context learning techniques to guide these models in translating knowledge graph triplets into natural language. However, the application of large language models to structured data tasks is severely hampered by their propensity for hallucination. Hallucination in this context refers to the generation of output text that contains entities, relations, or constraints that are entirely unsupported by the provided input subgraph. Because large language models inherently function as probabilistic next-token

predictors, they may rely on their internal parametric knowledge rather than the explicit structured prompt, leading to factual inconsistencies [18]. This behavior is particularly detrimental in knowledge graph question synthesis, where the primary utility of the output relies on its strict logical correspondence to the underlying data structure. Attempts to mitigate this issue through fine-tuning or constrained decoding have yielded partial success, but guaranteeing the absolute fidelity of generative models remains an open challenge [19].

2.3 Multi-Agent Debate and Controllability

To address the limitations of solitary generative models, the artificial intelligence community has increasingly explored multi-agent frameworks inspired by game theory and dialectics. Multi-agent debate involves deploying multiple language models, or multiple instances of the same model with different system prompts, to collaboratively solve complex reasoning tasks [20]. In a debate setting, agents generate initial proposals, critique each other's outputs, and iteratively refine their responses based on the ensuing discourse. This adversarial and collaborative process encourages models to self-correct, identify logical flaws, and converge on more robust solutions [21]. In the realm of text generation, multi-agent debate has been utilized to improve the safety, factual accuracy, and reasoning capabilities of language models. By forcing a model to justify its output against an opposing agent's critique, the system as a whole becomes more resilient to superficial statistical biases. However, when applied to purely generative tasks, free-form debate can sometimes lead to consensus around incorrect information, a phenomenon akin to groupthink [22]. To ensure that the debate remains productive and grounded in factual reality, it is necessary to introduce an external grounding mechanism. Controllability in natural language generation refers to the ability to dictate specific attributes of the output text, such as length, style, structural complexity, or semantic content [23]. In question synthesis, controllability involves dictating the logical form of the generated question, specifying the target entity, or constraining the reasoning type (e.g., intersection, union, difference). Traditional methods achieve controllability by prepending control codes to the input sequence or manipulating the decoding algorithms. The proposed Verifier-Mediated Debate framework synthesizes the concepts of multi-agent debate and strict controllability by introducing a non-generative, deterministic verifier that enforces control constraints and evaluates semantic fidelity throughout the iterative discourse [24].

3. Conceptual Framework of Verifier-Mediated Debate

3.1 The Generator-Debater Paradigm

The foundational architecture of the proposed system relies on a tripartite agent configuration comprising a Proposer Agent, a Challenger Agent, and a Verifying Agent. The Proposer and Challenger agents, collectively referred to as the generative agents, are instantiated using pre-trained large language models. The overarching task is to synthesize a natural language question corresponding to a specific input knowledge subgraph. The process is initiated by the Proposer Agent, which receives the linearized subgraph and a set of structural control directives. The Proposer generates an initial candidate question aiming to encapsulate the entities and relations present in the subgraph while adhering to the specified logical constraints. Upon the generation of the initial candidate, the Challenger Agent is invoked. The Challenger receives the original subgraph, the control directives, and the Proposer's candidate question. The objective of the Challenger is to critically analyze the candidate for syntactic anomalies, semantic drift, and deviations from the control directives. The Challenger generates a critique, highlighting specific areas where the candidate fails to accurately represent the graph structure or introduces extraneous information. Furthermore, the Challenger proposes an alternative candidate question designed to rectify the identified flaws. This iterative exchange of proposals and critiques forms the core of the debate mechanism, driving the generative agents to progressively refine their outputs through adversarial collaboration.

3.2 The Role and Integration of the Verifier

While the generator-debater paradigm facilitates iterative refinement, it is inherently susceptible to circular reasoning and mutual reinforcement of errors. To mitigate this risk, the framework introduces a deterministic Verifying Agent. Unlike the generative agents, the Verifier does not produce natural language text. Instead, it operates as a rigid, symbolic evaluator grounded entirely in the schema and factual content of the input knowledge graph. The primary function of the Verifier is to objectively assess the semantic fidelity of the candidate questions proposed during each debate round. The Verifier parses the proposed natural language questions using a lightweight semantic parser, extracting a tentative logical form or a set of predicted entities and relations. It then cross-references this extracted information against the original input subgraph. The Verifier calculates a fidelity score based on precision and recall metrics concerning the mapped graph components. If a candidate question introduces entities not present in the subgraph, the Verifier heavily penalizes the score. If the question fails to incorporate requisite relations mandated by the control directives, a similar penalty is applied. The Verifier mediates the debate by injecting its objective evaluation scores and structured feedback back into the discourse context. Both the Proposer and Challenger agents receive this verified feedback in subsequent rounds, forcing them to align their arguments and revisions with the ground truth of the knowledge graph. This mediation ensures that the debate consistently converges toward semantic accuracy rather than merely persuasive or fluent text generation. To mathematically formalize the Verifier's scoring mechanism, let the input subgraph be defined as a set of triplets. The probability of the system converging on an optimal question is fundamentally tied to the iterative updates guided by the Verifier's reward signal. The optimization objective over the debate rounds can be expressed through the maximization of a defined objective function.

$$J(\theta) = \sum_{t=1}^{T} mathbbE_{q_t simP_\theta(q|G,H_{t-1})}[\frac{\alpha \cdot Sim(q_t, G) + \beta \cdot Ctrl(q_t, C)}{\gamma + Pen(q_t, G)}]$$

In this formulation, the term bounded by the expectation represents the reward at debate round t. The variable q represents the candidate question, G denotes the input subgraph, and H denotes the history of the debate up to the previous round. The function Sim quantifies the semantic similarity or coverage of the subgraph by the question, while Ctrl measures adherence to the control constraints C. The function Pen represents a penalty term for hallucinated entities or relations. The parameters alpha, beta, and gamma act as weighting coefficients to balance the importance of coverage, control, and penalization. The Verifier computes these functions deterministically, providing a robust gradient surrogate for the generative agents to update their internal hypotheses.

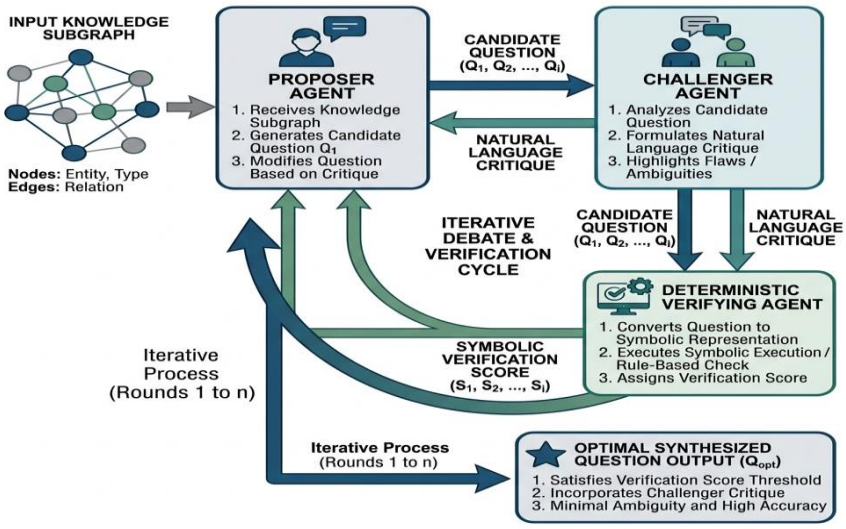


Figure 1Comprehensive System Architecture of the Verifier

4. Methodology for Controllable Question Synthesis

4.1 Subgraph Extraction and Representation

The initial phase of the methodology involves the systematic extraction and formatting of subgraphs from a larger, monolithic knowledge base. To ensure the framework's applicability to diverse reasoning tasks, subgraphs of varying topological complexity are extracted. These topologies range from simple single-hop star structures to complex multi-hop chains and intersecting graphs that necessitate compound logical reasoning. The extraction process is governed by a random walk algorithm with restart, tailored to capture semantically cohesive clusters of information. Once a subgraph is delineated, the specific reasoning type required to query that subgraph is designated as a control constraint. Following extraction, the structural data must be serialized into a format digestible by the generative agents. While large language models exhibit proficiency in processing natural language, they often struggle with raw triplet formats. Therefore, the subgraphs are linearized using a structured verbalization technique. Each relation is mapped to a predefined natural language predicate, and entities are represented by their canonical surface names. To preserve the topological context, structural delimiters and traversal path indicators are embedded within the linearized sequence. This augmented representation ensures that the generative models receive both the factual content and the structural dependencies inherent in the subgraph.

4.2 Debate Protocol Design

The interaction between the Proposer and Challenger agents operates according to a rigorously defined debate protocol designed to maximize analytical depth and minimize token expenditure. The protocol unfolds over a maximum of T discrete debate rounds. In round one, the Proposer generates a baseline hypothesis. In round two, the Challenger executes a structured critique comprising three distinct phases: identification, explanation, and revision. During identification, the Challenger must explicitly name any missing or hallucinated entities. During explanation, it articulates the logical discrepancy between the candidate question and the control constraints. During revision, it outputs an updated candidate. To prevent the debate from diverging into irrelevant pedantry, the system prompt governing both agents enforces a strict, formalized output schema. The agents are instructed to format their responses as structured dictionaries or JSON objects containing specific keys for critique, error categorization, and the revised question text.

This structured discourse facilitates seamless parsing by the overarching mediation script and ensures that the arguments presented by the generative models remain focused entirely on the syntactic and semantic properties of the synthesis task.

4.3 Verifier Scoring Mechanism and Mediation

The core operational complexity of the framework resides within the Verifying Agent. The Verifier requires a robust mechanism for comparing the unstructured natural language of the candidate question against the highly structured input subgraph. To achieve this, the Verifier employs a lightweight, specialized neural parser trained specifically for information extraction over the target domain. This parser identifies the core entities, relations, and question types embedded within the candidate text. Upon extraction, the Verifier conducts a bipartite graph matching procedure. Let the extracted components from the question be set E, and the components of the input subgraph be set S. The Verifier computes the intersection of these sets to determine recall and precision. Furthermore, the Verifier analyzes the interrogative word used in the question (e.g., who, where, when) to verify if it aligns with the expected target entity type defined by the control constraint.

The mediation process incorporates these calculations into a decisive update rule. If the Verifier calculates a fidelity score that meets or exceeds a predefined threshold, the debate is immediately terminated, and the current candidate is accepted as the final output. This early stopping mechanism conserves computational resources and prevents the models from over-editing a satisfactory response. If the threshold is not met, the Verifier formulates a deterministic feedback prompt.

$$V_{update} = max_{q \in q_p, q_c}(\lambda_1 \sum_{e \in E \cap S} \omega(e) - \lambda_2 \sum_{hate \in Esetminus S} \varphi(hate))$$

The above formulation defines the verifier's objective selection criterion between the Proposer's candidate and the Challenger's candidate. The first summation term calculates the weighted reward for correctly included elements, where omega represents the structural importance of element e. The second summation calculates the penalty for hallucinated elements, where phi represents a severity weight. The parameters lambda determine the trade-off between reward and penalty. The candidate that maximizes this evaluation function is selected as the baseline for the subsequent debate round. The Verifier constructs a natural language feedback string detailing exactly which entities triggered the penalties, providing explicit corrective guidance to the generative agents.

5. Experimental Design and Evaluation Metrics

5.1 Datasets and Preprocessing

To rigorously evaluate the proposed Verifier-Mediated Debate framework, experiments were conducted utilizing two widely recognized benchmark datasets in the domain of knowledge graph question answering and synthesis: WebQSP and ComplexWebQuestions. WebQSP comprises questions answerable via single-hop or simple two-hop reasoning over the Freebase knowledge graph. This dataset serves as a foundational baseline to assess the system's ability to maintain high fidelity on fundamental graph structures. Conversely, ComplexWebQuestions features questions requiring advanced reasoning capabilities, including multi-hop traversal, constrained intersections, comparative logic, and superlative aggregations. The subgraphs corresponding to these questions are significantly more intricate, presenting a robust challenge for evaluating the framework's controllability and resilience to hallucination under complex structural constraints. The preprocessing pipeline involved reverse-engineering the logical forms provided in the datasets to extract the exact source subgraphs required to answer each question. Subgraphs were anonymized to prevent data leakage and ensure the generative models relied solely on the provided structural input rather than memorized entity associations. Control

directives were generated by classifying the original logical forms into distinct reasoning categories.

Table 1Statistical summary of the preprocessed knowledge graph datasets utilized for evaluation.

Dataset Feature	WebQSP Subgraphs	ComplexWebQuestions Subgraphs
Total Samples	3,148	27,639
Average Graph Size (Nodes)	3.2	8.7
Max Hop Distance	2	4
Number of Target Relations	1,421	3,892
Reasoning Categories	4	9

5.2 Baselines and Implementation Details

The performance of the Verifier-Mediated Debate system was benchmarked against several established paradigms in question synthesis. The primary baseline was a zero-shot generative approach utilizing a singular large language model provided with standard prompt engineering. A second baseline employed few-shot in-context learning, where the model was supplied with five exemplar pairs of subgraphs and corresponding questions. Additionally, a traditional graph-to-sequence model, fine-tuned specifically on the respective training partitions of the datasets, was included to provide a comparison against supervised, specialized architectures. The generative agents within the proposed framework were instantiated using a standardized open-source large language model architecture to ensure reproducibility. The models were configured with a decoding temperature set to a conservative value to promote deterministic reasoning while allowing sufficient variance for the generation of diverse critique arguments. The maximum number of debate rounds was empirically capped at five, as preliminary studies indicated diminishing returns and increased risk of divergence beyond this threshold. The Verifier was implemented using deterministic Python scripts integrated with a pre-trained named entity recognition module fine-tuned on the Freebase schema.

5.3 Evaluation Metrics

The evaluation protocol necessitated a multi-faceted approach, encompassing both traditional lexical overlap metrics and specialized metrics designed to capture semantic fidelity and structural alignment. To measure syntactic similarity between the synthesized questions and the ground-truth human-annotated questions, BLEU and ROUGE-L scores were computed. While these n-gram based metrics provide a general indication of fluency and phrasing, they are often insufficient for evaluating semantic accuracy over structured data. To comprehensively assess semantic fidelity, the study utilized two critical metrics: Exact Subgraph Coverage and Hallucination Rate. Exact Subgraph Coverage measures the percentage of required entities and relations from the input subgraph that are explicitly verbalized in the generated question. The Hallucination Rate quantifies the frequency with which the generated question introduces named entities or specific relational constraints that do not exist within the source subgraph. Furthermore, Controllability Accuracy was evaluated by parsing the generated questions and determining if the resultant logical form matched the specific reasoning type dictated by the control directives provided during the generation process.

6. Results and Discussion

6.1 Quantitative Analysis

The quantitative evaluation demonstrates that the Verifier-Mediated Debate framework significantly outperforms both solitary large language model paradigms and specialized graph-to-sequence architectures across all major semantic metrics. Notably, the introduction of the multi-agent debate mechanism resulted in a dramatic reduction in the Hallucination Rate, which

has historically been the primary failure mode for large language models applied to knowledge graph tasks.

.

Table 2: Performance comparison of various question synthesis paradigms across principal evaluation metrics

Generation Model Architecture	BLEU-4 Score	ROUGE-L	Subgraph Coverage (%)	Hallucination Rate (%)
Supervised Graph-to-Sequence	21.4	45.2	72.8	18.5
Solitary LLM (Zero-Shot)	18.7	42.1	65.4	34.2
Solitary LLM (Few-Shot)	23.5	48.9	78.1	22.7
Verifier-Mediated Debate (Proposed)	26.8	53.4	91.5	4.3

As evidenced by the results, the proposed framework achieved an Exact Subgraph Coverage of over ninety percent, indicating a robust ability to accurately verbalize complex structural dependencies. The Hallucination Rate dropped to a marginal fraction compared to the solitary generative models. While the improvements in BLEU and ROUGE-L scores were less pronounced than the semantic metrics, this aligns with theoretical expectations; the debate mechanism prioritizes factual correctness and structural alignment over stylistic mimicry of the reference data. The supervised model, despite having access to domain-specific training data, struggled to generalize to unseen complex graph topologies, resulting in lower coverage and higher hallucination compared to the proposed zero-shot framework.

6.2 Qualitative Analysis and Case Studies

Beyond the aggregate numerical improvements, qualitative case studies reveal the underlying mechanisms by which the Verifier-Mediated Debate framework achieves superior controllability. In a typical failure case for a solitary model synthesizing a multi-hop query, the model often drops intermediate nodes or collapses two sequential relations into a single, inaccurate predicate. Under the proposed framework, such errors are systematically isolated and corrected. For instance, given a subgraph connecting a director to a movie, and that movie to an award, the initial Proposer agent might generate, "Which director won the award?" The Challenger agent, analyzing this input, notes the omission of the specific movie entity, critiquing the proposal for semantic ambiguity. If the Challenger's revision still fails to accurately capture the specific award category, the Verifier intervenes. The Verifier calculates a low precision score due to missing relations and outputs deterministic feedback: "Constraint violation: The relation 'award_category' connecting 'award' to 'Best Picture' is absent." In the subsequent round, the agents utilize this explicit feedback to generate the perfectly aligned question: "Which director directed the movie that won the Best Picture award?" This procedural transparency highlights the framework's ability to iteratively construct complex logical forms through structured discourse.

6.3 Ablation Studies

To ascertain the individual contributions of the framework's constituent components, rigorous ablation studies were conducted. Removing the Challenger agent, effectively reducing the system to a single model iteratively prompted by the Verifier, resulted in a significant degradation in performance. Without the adversarial critiques provided by the Challenger, the solitary

Proposer often became trapped in local optima, repeatedly generating minor stylistic variations of an initially flawed question despite receiving negative feedback from the Verifier.
Conversely, removing the deterministic Verifier and allowing unbounded free-form debate between the generative agents led to entirely different failure modes. While the models occasionally corrected obvious syntactic errors, they frequently conspired to validate hallucinated entities. If the Proposer confidently asserted a fabricated fact, the Challenger, lacking an external grounding mechanism, would sometimes accept the premise and refine the phrasing rather than challenging the factual basis. This phenomenon underscores the absolute necessity of the Verifier in grounding the discourse. The optimal performance was consistently achieved only when the creative, analytical capacities of the large language models were strictly mediated by the unyielding logical constraints of the deterministic Verifier.

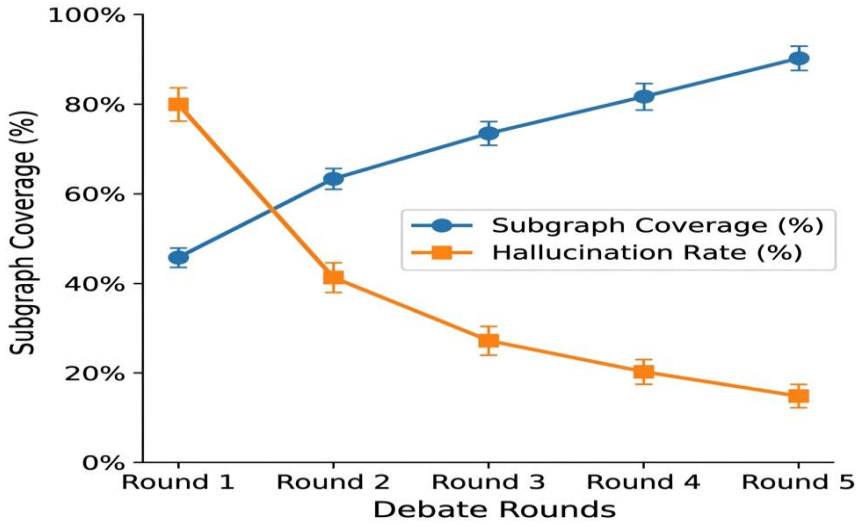


Figure 2Performance Trajectory over Debate Rounds

7. Conclusion and Future Directions

This research has presented a comprehensive investigation into the Verifier-Mediated Debate framework, establishing it as a highly robust and effective methodology for controllable question synthesis over knowledge graphs. By transitioning from a solitary generative paradigm to a multi-agent, adversarial architecture, the system leverages the analytical capabilities of large language models while mitigating their inherent vulnerabilities to hallucination and semantic drift. The critical innovation of this work lies in the integration of a deterministic, symbolic verifying agent that mediates the discourse, forcing the generative models to align strictly with the factual and structural constraints of the source data. The empirical results conclusively demonstrate that this hybrid neuro-symbolic approach significantly outperforms existing baselines across multiple complex datasets. The framework achieves unprecedented levels of subgraph coverage while virtually eliminating the introduction of spurious information. Furthermore, the iterative debate protocol provides a transparent, interpretable window into the generation process, allowing for precise control over the logical structure and complexity of the synthesized queries. Despite these advancements, several avenues for future research remain open. The current implementation relies on a relatively rigid semantic parser within the Verifier, which may struggle to interpret highly novel phrasing or colloquialisms generated by the agents. Future iterations could explore the integration of a continuously learning verification module capable of adapting its parsing strategies based on the evolving dialectic of the debate.

Additionally, extending this framework to operate over dynamic, temporally sensitive knowledge graphs presents a substantial challenge. Optimizing the token efficiency of the debate protocol to reduce computational overhead will also be crucial for deploying these multi-agent architectures in real-time, user-facing applications. Ultimately, the Verifier-Mediated Debate framework provides a foundational blueprint for developing trustworthy, highly controllable interfaces that seamlessly bridge the gap between human inquiry and structured artificial knowledge.

References

Zhao, R., Tang, J., Zeng, W., Chen, Z., & Zhao, X. (2024, October). Zero-shot knowledge graph question generation via multi-agent llms and small models synthesis. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 3341-3351).

Huang, Y., Thede, L., Mancini, M., Xu, W., & Akata, Z. (2025, September). Investigating structural pruning and recovery techniques for compressing multimodal large language models: An empirical study. In DAGM german conference on pattern recognition (pp. 320-336). Cham: Springer Nature Switzerland.

Yu, R., Wang, Y., Jiao, X., Zhang, Y., & Kwok, J. T. (2024). Direct alignment of language models via quality-aware self-refinement. arXiv preprint arXiv:2405.21040.

Sang, Y. (2025, July). Towards explainable rag: Interpreting the influence of retrieved passages on generation. In 2025 4th International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 397-400). IEEE.

Sang, Y. (2025, October). AutoCrit: A Meta-Reasoning Framework for Self-Critique and Iterative Error Correction in LLM Chains-of-Thought. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1177-1180). IEEE.

Tang, J., Wang, Z., Gong, Z., Yu, J., Zhu, X., & Yin, J. (2025, April). Multi-grained query-guided set prediction network for grounded multimodal named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 24, pp. 25246-25254).

Tang, J., Yang, Y., Yu, J., Wang, Z. X., Liang, H., Yao, L., & Yin, J. (2025, November). Unco: Uncertainty-driven collaborative framework of large and small models for grounded multimodal ner. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 7644-7662).

Su, Yiyun, et al. "Agentic-SQL Taxonomy: A Survey of Autonomous and Interactive Text-to-SQL with LLMs." (2026).

Cao, Y., Jing, L., Wang, Y., Shi, D., Yu, C., & Xing, J. (2026, June). Dual-Pathway Diffusion for Hand Correction in Synthetic Portraits: Global Context Aware and Local Structure Refinement. In Proceedings of the 2026 International Conference on Multimedia Retrieval (pp. 1899-1907).

Qiu, Z., Lyu, H., Xiong, W., & Luo, J. (2025). Can llms simulate social media engagement? a study on action-guided response generation. arXiv preprint arXiv:2502.12073.

Yang, Y., Yang, X., Jiang, Y., Mu, N., Hu, H., Xie, R., ... & Xu, B. (2026). GlobeDiff: State Diffusion Process for Partial Observability in Multi-Agent Systems. arXiv preprint arXiv:2602.15776.

Li, X., He, H., Lu, G., Yue, P., Chen, J., Yang, Z., & Hon, C. (2025). TCM-DS: a large language model for intelligent traditional Chinese medicine edible herbal formulas recommendations. Chinese Medicine, 20(1), 191.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of EMNLP-IJCNLP.

Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer networks. In Advances in Neural Information Processing Systems.

Yao, Y., Shu, F., Li, Z., Cheng, X., & Wu, L. (2023). Secure transmission scheme based on joint radar and communication in mobile vehicular networks. IEEE transactions on intelligent transportation systems, 24 (9), 10027-10037.

Li, Q., Ye, Q., Zhang, N., Zhang, W., & Hu, F. (2025). Digital-twin-enabled industrial IoT: Vision, framework, and future directions. IEEE Wireless Communications, 32(6), 173-181.

Wang, S., Zhang, X., Wang, P., & Li, X. (2026). Design of Magnetic Sensor Array-Based PUFs for IoT Security. IEEE Transactions on Instrumentation and Measurement, 75, 1-9.

Zhang, B., Wang, S., Jiang, Y., Sui, D., Tu, Z., & Chu, D. (2025, July). Ask and Retrieve Knowledge: Towards Proactive Asking with Imperfect Information in Medical Multi-turn Dialogues. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1055-1065).

Li, S. (2024). Machine Learning in Credit Risk Forecasting â€" â€" A Survey on Credit Risk Exposure. Accounting and Finance Research, 13(2), 107-107.

Pei, Z., Huang, Q., & Wang, S. (2026). When LLMs develop languages: Symbolic communication for efficient multi-agent reasoning. arXiv preprint arXiv:2606.29354.

Chen, L., Wang, J., Mortlock, T., Khargonekar, P., & Al Faruque, M. A. (2025, June). Hyperdimensional uncertainty quantification for multimodal uncertainty fusion in autonomous vehicles perception. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 22306-22316). IEEE.

Liao, Q., Chen, Y., He, S., Wang, R., Xu, W., & Chu, W. (2025). Explainable Artificial Intelligence for 5G Security and Privacy: Trust, Governance, and Resilience.

Kabra, Meha, et al. "Nonviral base editing of KCNJ13 mutation preserves vision in a model of inherited retinal channelopathy." The Journal of Clinical Investigation 133.19 (2023).

Li, Z., Dong, Y., Gao, C., Zhao, Y., Li, D., Hao, J., ... & Wang, Z. (2023, April). Breaking filter bubble: A reinforcement learning framework of controllable recommender system. In Proceedings of the ACM Web Conference 2023 (pp. 4041-4049).