Reflection-Driven Educational Knowledge Graph Expansion Method

Xin Zhang[a*], Yu Bai[b*] and Guiping Zhang

Shenyang Aerospace University, Shenyang 110000, Liaoning, China

[a]zx_1556822813@163.com, [b]baiyu@sau.edu.cn

Abstract. Construction of domain knowledge graphs is critical for intelligent learning and cognitive assessment, yet real-world expansion is often hindered by multi-source noise, leading to semantic drift and structural inconsistency. To address this, we propose RoKG-Agent, a reflection-driven framework for noisy term repair and sibling concept expansion. The method normalizes perturbed terms as semantic anchors and employs multi-dimensional evaluation with iterative optimization to ensure generation quality. We further introduce RoKG-Bench, a hierarchical noise benchmark derived from real-world teaching data. Experiments across multiple foundation models demonstrate that RoKG-Agent significantly improves robustness, structural accuracy, and semantic consistency under severe noise conditions.

Keywords: Domain Knowledge Graph, Large Language Model (LLM), Reflection-Driven Generation, Multidimensional Quality Assessment, Robustness

1 Introduction

Knowledge Graph Construction (KGC) aims to transform unstructured or semi-structured data into structured knowledge representations[1]. In vertical domains, educational knowledge graphs model disciplinary concepts and their logical relations, providing essential support for applications such as intelligent tutoring and cognitive diagnosis. Compared with general-purpose knowledge graphs, educational scenarios impose stricter requirements on data quality, with extremely low tolerance for knowledge deviation and incorrect relations. Even minor errors in concept hierarchies may mislead learners and undermine the consistency of the instructional system. In practical dynamic expansion settings, the input to educational knowledge graphs is often not a well-structured graph, but a collection of discrete, semi-structured data objects derived from heterogeneous sources, including OCR-scanned teaching materials[2], classroom speech transcripts[3], and semi-structured logs[4]. Such data typically contain various types of noise, such as character confusion, term truncation, and format perturbation. Existing KGC methods mainly focus on joint entity–relation extraction from clean corpora (e.g., PRGC[5], OneRel[6]) and tend to suffer from semantic boundary ambiguity and error propagation under noisy conditions.

I. A. Khan et al. (eds.), *Proceedings of the 2026 5th International Conference on Educational Innovation and Multimedia Technology (EIMT 2026)*, Atlantis Highlights in Social Sciences, Education and Humanities 51,
https://doi.org/10.2991/978-94-6239-691-3_60

Meanwhile, although large language models (LLMs) have shown strong generative capabilities, and recent studies have introduced self-correction strategies such as Chain-of-Thought (CoT)[7] and Reflexion[8], these approaches generally perform coarse-grained global refinement on single-pass outputs. They lack fine-grained diagnosis and iterative optimization mechanisms for distorted educational terminology. Under severe noise, LLMs are still prone to knowledge hallucination[9], which further leads to semantic drift in expanded concept nodes.

To address these challenges, this paper defines the task as educational terminology repair and sibling-level concept expansion under noisy conditions. We propose a reflection-driven robust knowledge graph expansion framework (RoKG-Agent). The proposed framework first normalizes and repairs noisy terms as semantic anchors, and then generates candidate sibling concepts. A multi-dimensional evaluation mechanism is introduced to jointly constrain the outputs, and an iterative refinement strategy is triggered based on predefined thresholds. Finally, high-confidence knowledge is integrated into the knowledge graph.

2 Methods

2.1 Task Definition and Overall Architecture

This work focuses on knowledge graph expansion under noisy input, involving two consecutive subtasks. Given an input text sequence containing multi-source noise with perturbed entities, the objectives are defined as follows. (1) Terminology Repair: We filter out noise to reconstruct standardized terms with clear boundaries , formalized as $E_{norm} = f_{repair}(T_{noisy})$ and (2) Peer-Level Concept Expansion: Using the repaired term as a semantic anchor, we generate a set of candidate nodes belonging to the same ontological level without strict external knowledge base constraints, enabling horizontal expansion: $E_{sib} = \{e_1, e_2, \ldots, e_k\} = f_{expand}(E_{norm})$. Here, k denotes the number of peer-level concepts generated, and each node in the set E_{sib} must be consistent with e_i and E_{norm} in terms of logical and semantic categories. The proposed RoKG-Agent framework is designed to address these two subtasks.

To address these subtasks, the proposed RoKG-Agent framework (Fig. 1) operates on a "Repair—Expansion—Evaluation—Reflection" closed-loop paradigm to process discrete, semi-structured input data. It incrementally updates the knowledge graph without external gold-standard supervision through three stages: (1) Repair & Expansion: Normalizes noisy entities into standard anchors to generate peer-level candidates; (2) Multidimensional Assessment: Evaluates outputs against predefined thresholds across normativity, fidelity, homogeneity, and consistency; and (3) Iterative Reflection: Uses diagnostic feedback to trigger targeted regeneration until integration standards are met.

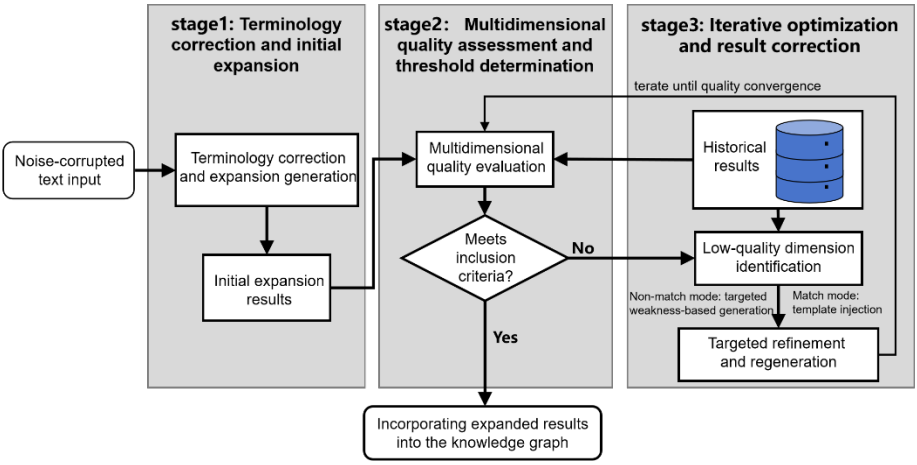

Fig. 1. RoKG-Agent framework diagram

2.2 Multidimensional Fine-Grained Evaluation and Dual-Threshold Decision

To mitigate semantic drift and structural errors during generative expansion, we replace traditional single-score judgments with a multidimensional evaluation module. It quantifies generation quality across four independent dimensions: (1) Terminology Normativity: Penalizes character and formatting anomalies to assess term standardization (Eq. 3); (2) Semantic Fidelity: Measures semantic consistency between the repaired term and the original noisy text via Sentence-BERT[10] cosine similarity (Eq. 4); (3) Structural Homogeneity: Computes the average pairwise semantic similarity among candidate nodes to prevent category contamination (Eq. 5); and (4) Logical Consistency: Verifies whether the concept category implied by the repaired term aligns with its generated definition (Eq. 6).

After computation, the system integrates these metrics into a weighted comprehensive score (Eq. 7). A dual-threshold mechanism (Eq. 8) is then applied: results are integrated into the knowledge graph only when all individual dimension scores and the comprehensive score exceed their respective thresholds. Otherwise, the lowest-scoring dimension triggers targeted optimization prompts for iterative regeneration, effectively suppressing noise propagation and ensuring overall expansion reliability.

$$S_{norm}(E) = 1 - \alpha \cdot \frac{\sum_{t \in T} (I(t \in C) + I(t \in W))}{|T|} \tag{1}$$

$$S_{fid}(E, D) = \mathrm{CosSim}(\mathbf{v}_E, \mathbf{v}_D) \tag{2}$$

$$S_{hom}(N_{sib}) = AvgSim(N_{sib}) \tag{3}$$

$$S_{alg}(N, D) = f_{logic}(\mathrm{C}(N) \leftrightarrow \mathrm{C}(D)) \tag{4}$$

$$S_{diag} = \omega \cdot [S_{norm}, S_{fid}, S_{hom}, S_{alg}]^T \tag{5}$$

$$\text{Condition} = (\forall S_i > \tau_i) \wedge (S_{diag} > T) \tag{6}$$

2.3 Trajectory Recording and Dual-Channel Memory Coordination Mechanism

To enhance multi-round iteration stability under complex noise, RoKG-Agent incorporates a dual-channel memory coordination mechanism to provide stable contextual constraints. It consists of: (1) Historical Trajectory, which structurally stores noisy inputs (processed as discrete data objects), intermediate repaired terms, and evaluation vectors in each cycle to effectively prevent repeated errors and semantic drift; and (2) Pattern Memory, which predefines repair strategies for high-frequency structural noise (e.g., symbol fragmentation) Crucially, this memory acts as a "fast-path" cache that bypasses redundant LLM computations, significantly reducing inference latency and computational costs in large-scale applications. It is triggered on-demand by low-score diagnostic signals. During the iterative phase, the system coordinates both memory channels based on evaluation feedback to generate the next optimization instruction:

$$P_{next} = \begin{cases} \mathrm{T}(\sigma) + \mathrm{H}(\Gamma), & \text{if Pattern Hit} \\ \mathrm{A}(S_{\min}) + \mathrm{H}(\Gamma), & \text{if Pattern Miss} \end{cases} \tag{7}$$

3 Results and Analysis

To evaluate the robustness and stability of the proposed framework under complex noise, we conduct experiments centered on gradually increasing noise intensity. Our method is compared against representative reasoning- and reflection-based approaches across LLMs of varying scales, followed by ablation studies to validate key components. To simulate real-world industrial scenarios, we first construct a hierarchical benchmark, RoKG-Bench, derived from a computer science teaching knowledge graph. By applying automated noise to instructor-verified teaching concepts, we establish an objective ground truth that structurally eliminates the need for subjective human annotation and inter-annotator agreement (IAA) checks. Mirroring real-world discrete inputs, each base sample is defined as an independent JSON object capturing the local knowledge topology. We select 300 core terms as seed nodes and apply an automated noise injection strategy to generate 900 test samples (three noisy variants per node) categorized into three intensities:

Level-1 (Symbolic Noise): Light formatting disturbances (e.g., random spaces, broken characters) to assess basic term standardization.

Level-2 (Structural Noise): Character-level interference (e.g., homophones, visually similar substitutions) to evaluate structural and semantic recovery.

Level-3 (Semantic Composite Noise): Complex disturbances (e.g., irrelevant text insertion, fragmentation) to test robustness against severe semantic drift.

3.1 Results Analysis and Discussion

Table 1. Comparison results of different models on RoKG-Bench

Foundation model	Method	Level-1		Level-2		Level-3	
		F1	ASS	F1	ASS	F1	ASS
Qwen2.5-3B	CoT Prompt	93.06	91.52	88.91	87.55	81.51	82.03
	Reflexion	94.84	**95.85**	91.91	90.54	83.31	84.56
	Ours	**96.30**	94.21	**92.73**	**94.12**	**86.47**	**91.28**
DeepSeek-R1-8B	CoT Prompt	94.67	92.88	89.48	89.50	83.13	85.12
	Reflexion	95.23	94.56	**93.38**	91.88	85.69	87.25
	Ours	**97.49**	**96.92**	92.82	**93.54**	**87.96**	**93.10**
GLM-4-9B	CoT Prompt	95.78	93.55	90.98	90.21	85.56	86.54
	Reflexion	97.00	95.10	92.23	92.56	87.53	88.42
	Ours	**98.67**	**97.83**	**93.73**	**96.25**	**89.34**	**94.58**

Table 1 reports performance on RoKG-Bench, including CoT Prompt as a representative single-pass baseline. While all methods perform well under Level-1 and Level-2 noise, the single-pass approach drops significantly at Level-3. In contrast, our method maintains stable structural accuracy and semantic similarity across backbone models. This performance gap under severe noise justifies the computational overhead of the agent's iterative reflection.

Under Level-2 with DeepSeek-R1, Reflexion achieves a slightly higher F1 but much lower semantic similarity, indicating potential semantic drift from single-feedback reliance. Our multi-dimensional evaluation effectively preserves semantic consistency.

4 Conclusion

This paper proposes RoKG-Agent, a reflection-driven framework for robust knowledge graph expansion under noisy conditions. To mitigate term distortion and semantic drift, our approach replaces conventional single-pass generation with a closed-loop iterative optimization paradigm, jointly constrained by multidimensional quality evaluation and a dual-channel memory mechanism. Experiments on the RoKG-Bench benchmark demonstrate that RoKG-Agent achieves strong robustness across multiple LLMs, substantially improving structural accuracy and semantic consistency, especially under severe composite noise. Future work will focus on reducing inference latency via dynamic thresholding and extending the framework to complex tasks like joint entity-relation extraction. Furthermore, the domain-agnostic nature of the multidimensional

evaluation metrics allows RoKG-Agent to be readily adapted to other strict vertical domains, such as medical and legal knowledge graph construction.

References

1. Ji S, Pan S, Cambria E, et al. A Survey on Knowledge Graphs: Representation, Acquisition, and Appli-cations [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(2): 494-514.
2. Zhang J, Zhang Q, Wang B, et al. OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation [J]. arXiv preprint arXiv:2412.02592, 2024.
3. Alderete J, Hui M K F, Mohan A. Evaluating ASR Robustness to Spontaneous Speech Errors: A Study of WhisperX Using a Speech Error Database [C]. Proc. Interspeech 2025, 2025: 1803-1807.
4. Sui Y, Wang Y, Niu D, et al. LogKG: Log Failure Diagnosis Through Knowledge Graph [J]. IEEE Transactions on Services Computing, 2023, 16(5): 3493-3507.
5. Zheng H, Wen R, Chen X, et al. PRGC: Potential Relation and Global Correspondence Based Joint Re-lational Triple Extraction [C]. Proceedings of the 59th Annual Meeting of the Association for Compu-tational Linguistics (ACL), 2021: 6225–6235.
6. Shang Y M, Huang H, Mao X, et al. OneRel: Joint Entity and Relation Extraction with One Module in One Step [C]. AAAI Conference on Artificial Intelligence, 2022, 36(10): 11285-11293.
7. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models [C]. Proceedings of the 36th International Conference on Neural Information Processing Sys-tems (NIPS '22), 2022: 24824–24837.
8. Shinn N, Cassano F, Gopinath A, et al. Reflexion: language agents with verbal reinforcement learning [C]. Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23), 2023: 8634–8652.
9. Huang L, Yu W, Ma W, et al. A Survey on Hallucination in Large Language Models: Principles, Tax-onomy, Challenges, and Open Questions [J]. ACM Transactions on Information Systems, 2025, 43(2): 42.
10. Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [C]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019: 3982–3992.