

KG-RLVR: Reinforcing LLM Reasoning with Knowledge Graph-Derived Process Rewards

Aoran Gan¹, Guoxuan Zhang¹, Ya Li³, Qiangbing Wang³,
Quan Liu³, Qi Liu^{1,2*}, Guoping Hu^{1,3}

¹State Key Laboratory of Cognitive Intelligence,
University of Science and Technology of China

²Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

³Artificial Intelligence Research Institute, iFLYTEK Co., Ltd

{gar,tomarin}@mail.ustc.edu.cn, {yali8,qbwang2,quanliu}@iflytek.com, qiliuql@ustc.edu.cn*, gphu@iflytek.com

Abstract

As large language models (LLMs) are increasingly adapted from general-purpose systems into domain-specific problem solvers, a central challenge is how to provide reliable process-level guidance for knowledge-intensive reasoning. High-quality reasoning supervision, such as annotated reasoning trajectories or externally generated rationales, is costly to obtain, difficult to verify, and hard to scale across domains. In this paper, we propose KG-RLVR, a lightweight framework that reframes KG-enhanced LLM post-training as verifiable reward construction from traceable structured evidence. KG-RLVR retrieves evidence from external knowledge graphs, consolidates noisy graph neighborhoods into representative evidence units, and converts these units into locally computable process rewards. The reward encourages the model to cover directly relevant evidence while selectively integrating less direct but potentially useful auxiliary evidence, thereby guiding reasoning traces without requiring externally authored process demonstrations. Experiments on knowledge-intensive medical and multi-domain QA benchmarks show that KG-RLVR consistently improves over vanilla RL, demonstrating that KG-derived process rewards provide practical and scalable supervision for reasoning-oriented LLM post-training.

1 Introduction

Research on large language models (LLMs) has recently shifted from scaling general capabilities during pre-training to shaping task- and domain-specific behavior during post-training (Wu, 2025; Ouyang et al., 2022; Shao et al., 2024; Rafailov et al., 2023; Zhang et al., 2022a). For knowledge-intensive tasks, final-answer supervision alone is often insufficient: models must learn to select, combine, and apply relevant knowledge during reasoning (Lightman et al., 2024; Wang et al., 2024; Ma

et al., 2025). Yet result-level feedback is often too sparse to shape this process, whereas high-quality reasoning supervision is both expensive to obtain and difficult to verify at scale (Luo et al., 2024a; Setlur et al., 2025; Peng et al., 2025). Together, these limitations pose a key challenge for scalable reasoning-oriented post-training.

Knowledge graphs (KGs) offer a scalable source of structured and traceable evidence for such process-level guidance, and have therefore been widely studied as a means of enhancing LLMs (Vanapalli et al., 2025; Ma et al., 2025; Lavrinovics et al., 2025). Existing LLM-KG research has largely followed two paradigms. One line of work uses KGs at inference time, retrieving graph evidence to augment the input or validate model outputs, as in HippoRAG (Gutiérrez et al., 2024), KAPING (Baek et al., 2023), KGR (Chen et al., 2024), GoG (Xu et al., 2024), and GIVE (He et al., 2024). Another line incorporates KG evidence during training to improve knowledge-grounded capabilities, for example through data augmentation, entity-centric supervision, concept grouping, or retrieval-oriented instruction generation, as in RLKGF (Yan et al., 2025), KG-SFT (Chen et al., 2025), SAKE (He et al., 2026), and KG-Reasoner (Wang and Yu, 2026). While these studies show that KG evidence can support LLMs either as external context at inference time or as auxiliary signals for training-time capability shaping, its use as a direct and verifiable source of fine-grained process supervision for knowledge-intensive QA post-training remains comparatively underexplored.

Among KG-enhanced post-training methods, KG-SFT (Chen et al., 2025) provides a particularly relevant reference point. It uses retrieved KG evidence to ground teacher-generated rationales and trains the model through supervised fine-tuning (SFT), demonstrating that KG evidence can be converted into useful training supervision. Building on this insight, we ask whether KG evidence can super-

*Corresponding author.

wise reasoning more directly during post-training: rather than first rewriting graph evidence into a fixed natural-language rationale, can structured evidence itself be transformed into reward signals that guide how the model uses knowledge in its own reasoning process?

Motivated by reinforcement learning with verifiable rewards (RLVR), we propose KG-RLVR, a novel framework for KG-guided RL post-training, which converts retrieved KG-derived evidence units into verifiable process rewards. Here, KG-derived evidence units refer to compact and traceable knowledge units distilled from retrieved KG triples and short paths. These units are then used as rule-verifiable targets for matching against the model’s generated reasoning trace.

The core of KG-RLVR is a KG-reward-guided RLVR optimization process that turns KG-derived evidence units into training-time signals for reasoning. Rather than explicitly specifying a pre-determined reasoning process, KG-RLVR enables the model to learn through policy optimization how to organize, combine, and apply KG-derived evidence in ways that support answer discrimination. The KG-guided reward realizes this objective by promoting coverage of directly relevant evidence while encouraging selective integration of less direct but potentially useful auxiliary evidence. In this way, KG-RLVR steers policy optimization toward reasoning traces that are more evidence-grounded and answer-supportive.

Viewed from this perspective, KG-RLVR points to a different way of providing process supervision for LLM post-training. Rather than relying on high-quality reasoning texts, annotated retrieval trajectories, or demonstrations of intermediate actions, KG-RLVR derives process-level signals from traceable KG evidence without externally authored process annotations. The model is not trained to imitate a prescribed reasoning path; instead, it is rewarded for producing reasoning traces that cover and selectively use relevant KG-derived evidence. This reframes KG-enhanced post-training as reward construction from verifiable structured evidence. The framework remains lightweight: it only requires reward-level integration during post-training and preserves the original inference protocol.

Our contributions are threefold:

- We propose KG-RLVR¹, a novel framework that reframes KG-enhanced post-training as

verifiable reward construction from traceable KG-derived evidence, rather than supervision from externally generated reasoning texts or annotated retrieval trajectories.

- We design a KG-guided reward that encourages evidence-grounded reasoning by balancing core-evidence coverage, selective evidence use, and final-answer correctness.
- We show that KG-RLVR consistently improves over vanilla RL on knowledge-intensive QA benchmarks, while requiring only reward-level integration and preserving the original inference protocol.

2 Methodology

We study domain-specific multiple-choice QA. Given a question stem q , candidate options $\mathcal{O}$, and a ground-truth label y^* , KG-RLVR improves the policy through RL post-training with KG-derived supervision. For each instance, we retrieve triples and short paths from external KGs and render them as evidence units. We then cluster these units to reduce redundancy, partition them according to their connection to the input, and assign weights for reward computation. The resulting evidence sets define locally verifiable process rewards. Below, we describe these two stages in detail.

2.1 KG Retrieval and Evidence Clustering

For each training instance, KG-RLVR performs unsupervised, option-aware, and label-agnostic KG retrieval. We perform named entity recognition (NER) and entity linking over the question stem and all candidate options using external KGs. The linked entities serve as seed nodes for KG retrieval. Starting from these seed nodes, we expand the KG neighborhood by two hops and retrieve candidate triples and short paths. Each retrieved triple or short path is rendered through its entity and relation labels, yielding a label-based textual rendering. We treat these label-based textual renderings as the observable realizations of KG-derived evidence units. Formally, all retrieved evidence units for an instance form a set $\mathcal{E}$. At this stage, however, $\mathcal{E}$ remains highly redundant, as neighboring KG structures often contain overlapping entities, repeated relations, and semantically similar evidence units.

To reduce this redundancy before reward construction, we perform text clustering over the label-based textual renderings in $\mathcal{E}$, using the same

¹<https://github.com/AaronGalan/kgrrlvr>

KG-RLVR Overview

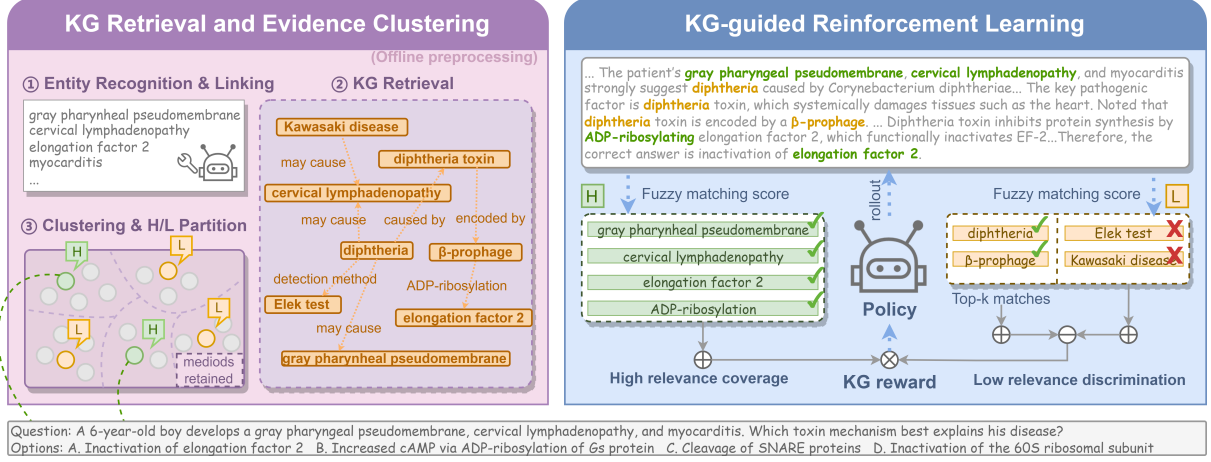


Figure 1: Overview of KG-RLVR. KG-RLVR performs offline KG retrieval and evidence clustering to obtain compact KG-derived evidence units, which are partitioned into a high-relevance set $\mathcal{H}$ and a low-relevance set $\mathcal{L}$. During RL post-training, fuzzy matching between the reasoning trace and the evidence units produces a KG-guided reward that encourages coverage of core evidence and selective use of auxiliary evidence, thereby promoting more evidence-grounded reasoning.

Levenshtein-based text similarity later adopted for RL reward matching. For each cluster, we retain the medoid and discard the remaining units, producing a condensed set $\mathcal{E}'$. The retained medoids provide representative evidence anchors while reducing sensitivity to duplicate and locally redundant graph content. This compact set is then used for reward construction.

Even after clustering, useful evidence may remain mixed with weakly related graph context. We therefore partition $\mathcal{E}'$ into a high-relevance set $\mathcal{H}$ and a low-relevance set $\mathcal{L}$. Units directly connected to the question or candidate options through seed entities enter $\mathcal{H}$; the remaining units enter $\mathcal{L}$ as less direct but potentially useful evidence. This partition allows the reward to distinguish core-evidence coverage from selective auxiliary-evidence use. Details of the clustering procedure and medoid selection are provided in Appendix A.

2.2 KG-guided Reinforcement Learning

The preceding stage produces a fixed high-relevance set $\mathcal{H}$ and low-relevance set $\mathcal{L}$ for each training instance. During RL post-training, these sets serve as reference evidence for evaluating the model's generated reasoning. Given an input $x = (q, \mathcal{O})$, the policy π_θ generates a response a consisting of a reasoning trace r and a final prediction $\hat{y}$. We treat both $\hat{y}$ and the ground-truth answer y^* as sets of selected option labels. The key design question is how $\mathcal{H}$ and $\mathcal{L}$ should jointly supervise

the generated reasoning trace r . The former should reward coverage of directly relevant KG evidence, while the latter should encourage selective use of auxiliary evidence and discourage indiscriminate copying.

KG-RLVR addresses this question by converting the two evidence sets into locally verifiable reward targets. Instead of requiring a complete reference reasoning trajectory, the method evaluates the model's own reasoning trace against KG-derived evidence units at the text level. The answer reward is determined by the final prediction $\hat{y}$, whereas the KG reward measures the interaction between r and the precomputed evidence sets $\mathcal{H}$ and $\mathcal{L}$. We next define these reward components in detail.

Specifically, we first measure how well the generated reasoning trace r matches each KG-derived evidence unit. For each evidence unit e , we use a Levenshtein-based fuzzy-matching function to locate its best textual match within r , yielding:

$$m(e, r) \in [0, 1],$$

where $m(e, r) = 1$ indicates an exact match. This text-level matching provides a lightweight and locally verifiable proxy for evidence usage in the reasoning trace. Based on these matching scores, we construct two complementary KG reward components: a high-relevance coverage term over $\mathcal{H}$ and a low-relevance discrimination term over $\mathcal{L}$.

For the high-relevance set $\mathcal{H}$, we define the coverage score as:

$$\bar{m}_{\mathcal{H}} = \frac{\sum_{e \in \mathcal{H}} w_e m(e, r)}{\sum_{e \in \mathcal{H}} w_e}.$$

The evidence weights $w_e > 0$ are derived from BM25-based ranking and group membership. This coverage term rewards reasoning traces that match high-relevance evidence units, thereby establishing a core-evidence anchor for KG-guided process supervision.

For the low-relevance set $\mathcal{L}$, the goal is not to maximize coverage over all of the units, but to encourage selective use of potentially useful auxiliary evidence. Let $\mathcal{L}_{\text{top}}$ denote the top $k_{\mathcal{L}}$ evidence units in $\mathcal{L}$ with the highest matching scores $m(e, r)$, and let $\mathcal{L}_{\text{rest}} = \mathcal{L} \setminus \mathcal{L}_{\text{top}}$. We define the low-relevance discrimination margin as:

$$\Delta \bar{m}_{\mathcal{L}} = \frac{\sum_{e \in \mathcal{L}_{\text{top}}} w_e m(e, r)}{\sum_{e \in \mathcal{L}_{\text{top}}} w_e} - \frac{\sum_{e \in \mathcal{L}_{\text{rest}}} w_e m(e, r)}{\sum_{e \in \mathcal{L}_{\text{rest}}} w_e}.$$

Here, $k_{\mathcal{L}}$ is the number of selected low-relevance evidence units, defined as:

$$k_{\mathcal{L}} = \lfloor r_k \frac{|\mathcal{L}|^2}{|\mathcal{L}| + |\mathcal{H}|} \rfloor$$

where $r_k \in (0, 1)$ controls the selection ratio. This term rewards the model for selectively using a subset of potentially useful low-relevance evidence units rather than copying all available content. Since $\mathcal{L}_{\text{top}}$ is selected according to the highest matching scores and all weights are non-negative, this margin is non-negative by construction. Accordingly, the KG reward is defined as:

$$R_{\text{KG}} = (\bar{m}_{\mathcal{H}})^{\lambda} \cdot (\Delta \bar{m}_{\mathcal{L}})^{1-\lambda}, \quad \lambda \in [0, 1]$$

where λ controls the trade-off between high-relevance evidence coverage and low-relevance evidence discrimination. By construction, we have $\bar{m}_{\mathcal{H}}, \Delta \bar{m}_{\mathcal{L}}, R_{\text{KG}} \in [0, 1]$.

In addition, KG-RLVR uses standard answer and format rewards. The answer reward is defined as the squared F1 score between the predicted label set and the ground-truth label set:

$$R_{\text{acc}} = F_1^2(\hat{y}, y^*)$$

where

$$F_1(\hat{y}, y^*) = \frac{2|\hat{y} \cap y^*|}{|\hat{y}| + |y^*|}$$

For single-choice questions, R_{acc} is equivalent to exact-match accuracy. For multi-choice questions, this reward provides partial credit and encourages the model to approach the correct answer set progressively.

The final reward is:

$$R = \alpha R_{\text{format}} + (1 - \alpha)(1 - \beta + \beta R_{\text{KG}}) R_{\text{acc}}, \\ \alpha, \beta \in [0, 1],$$

where R_{format} encourages the response to follow the required template. The KG reward modulates the answer reward and therefore contributes to task learning only when $R_{\text{acc}} > 0$. If $R_{\text{acc}} = 0$, matched evidence yields no non-format reward. Accordingly, R_{KG} serves as an outcome-conditioned process signal rather than as a standalone verifier of reasoning correctness. The coefficient α balances the format and task rewards, while β controls the strength of KG-based modulation. Table 3 further supports this gating design through comparison with a conventional weighted-sum formulation.

Within this gated objective, the partitioned evidence sets define complementary preferences over reasoning traces. The high-relevance coverage term favors trajectories that incorporate directly relevant evidence from $\mathcal{H}$, providing a core evidence anchor for reasoning. The low-relevance discrimination term favors concentrated matching to a small subset of potentially useful evidence from $\mathcal{L}$ over diffuse matching across the entire set, thereby discouraging indiscriminate evidence copying. Because these preferences are coupled with answer correctness, they are reinforced only when the resulting prediction receives nonzero answer credit. Together, the two components guide the policy toward reasoning traces that cover core evidence while selectively incorporating auxiliary evidence in an answer-supportive manner, without requiring externally authored rationales, teacher-generated reasoning traces, or annotated retrieval paths.

The resulting training procedure remains lightweight. KG retrieval, evidence clustering, partitioning, and BM25 weighting are all performed offline before RL training. During training, the pre-computed evidence units are matched against generated reasoning traces to produce the KG-guided scalar reward. In our experiments, we optimize the policy with GRPO and include entropy and KL-divergence auxiliary losses. The scalar reward is also compatible with other RL optimizers. At

inference time, the model follows the backbone’s direct-generation protocol and requires no KG retrieval, graph encoding, or teacher-model interaction. More RL details, such as the Levenshtein-based score, format reward, and normalized weighting rules, are provided in Appendix A.

3 Experiments and Analysis

3.1 Experimental Setup

Datasets and Knowledge Bases. Following KG-SFT, we evaluate KG-RLVR on knowledge-intensive multiple-choice QA tasks. MMedBench (Qiu et al., 2024) serves as our primary benchmark for main comparison and subsequent analyses, where we use all six language-specific subsets with UMLS (Bodenreider, 2004) as the external KG. The LLM first extracts candidate medical mentions, which are then linked by MetaMap (Aronson, 2001) to UMLS concept unique identifiers. To further examine whether the proposed reward design extends beyond medical QA, we additionally report main results on 10 domain QA subsets (Hendrycks et al., 2021) with Wikidata as the external KG, following KG-SFT. At test time, the model receives only the question stem and candidate options, and predictions are extracted through rule-based label matching. We report direct-inference accuracy as the primary metric. Details of the datasets are provided in Appendix B.

Model and Training. We use Qwen2.5-7B-Instruct (Qwen et al., 2025) as the main backbone model and adopt GRPO (Guo et al., 2025) as the policy optimization algorithm with the proposed KG-guided reward. To enable direct comparison with prior KG-enhanced baselines in the main results, we also include a Llama2-based (Touvron et al., 2023) KG-RLVR run under the corresponding baseline setting. Unless otherwise specified, ablation studies and analysis experiments are conducted on MMedBench with Qwen2.5-7B-Instruct. During inference, the model follows the same input-output protocol as the backbone LLM. Hyperparameters and prompt templates are provided in Appendix B.

Baselines. We introduce both training-free and training-based methods as the baselines. i) Vanilla LLM: direct inference with open-source weights; ii) ToG (Sun et al., 2024): inference enhanced by beam search over KG; iii) KGR (Chen et al., 2024): answer refinement via active exploration of the KG; iv) KAPING (Baek et al., 2023): appending the re-

trieved triples to the QA prompt; v) StructGPT (Jiang et al., 2023): iterative reading and reasoning over structured data with the IRR framework; vi) Vanilla SFT: direct fine-tuning of LLM on QA pairs; vii) KG-SFT (Chen et al., 2025): fine-tuning with GPT-generated rationales grounded in KG subgraphs; viii) Vanilla RL: fine-tuning of LLM with reinforcement learning using the normalized accuracy and format reward.

3.2 Main Results

Table 1 reports the results on MMedBench. Under the Llama2-based setting used by prior KG-enhanced baselines, KG-RLVR consistently improves over Vanilla RL across all six language-specific subsets and achieves comparable average performance to KG-SFT, despite mixed subset-level differences. With Qwen2.5-7B-Instruct as a stronger backbone, KG-RLVR consistently outperforms both Vanilla RL and KG-SFT across the six subsets, achieving an average accuracy of 71.86 compared with 66.49 and 68.76, respectively. These results show that the gains do not simply come from applying RL to QA pairs; rather, KG-derived process rewards provide useful supervision beyond answer-level RL alone and remain effective even when the backbone already has stronger initial capabilities.

As shown in Table 2, KG-RLVR also achieves competitive results on multi-domain QA. We further inspect all generated outputs and find that its format-decoding success rate exceeds 97% across all evaluated subsets. We provide several output cases in Appendix D. For a focused analysis, we conduct subsequent ablation and analysis studies on MMedBench, given its role as our primary benchmark and its multilingual coverage.

3.3 Ablation Study

We study six ablated variants of KG-RLVR:

- **Prompt-appended evidence only.** Retrieved evidence units are appended to the prompt without an explicit KG reward, testing whether the model can exploit KG evidence merely as additional context.
- **No clustering.** All retrieved evidence units are directly used in the KG reward, testing the effect of removing evidence denoising.
- **Random evidence units.** Instance-relevant evidence is replaced with randomly sampled

Table 1: Comparison of the accuracy (%) of different methods on MMedBench. Unless otherwise indicated (†), all other groups adopt Llama2-7B-chat as the backbone. The best results achieved on Qwen2.5-7B-Instruct and Llama2-7B-chat are highlighted in bold and underlined, respectively.

Method	#Train	English	Chinese	Russian	Spanish	French	Japanese	Avg.
Vanilla LLM	-	28.20	28.37	51.17	32.97	12.76	11.10	27.43
ToG	-	34.27	28.13	48.42	35.59	12.47	19.61	29.75
KGR	-	33.15	26.88	47.52	34.74	13.39	17.29	28.83
KAPING	-	36.39	27.24	54.66	34.98	11.54	15.91	30.12
StructGPT	-	35.16	24.50	55.32	36.16	14.24	20.01	30.90
SFT	✓	47.80	65.02	65.62	43.44	27.37	30.16	46.56
KG-SFT	✓	49.25	<u>67.86</u>	<u>68.75</u>	<u>46.49</u>	33.51	32.66	49.75
Vanilla RL	✓	48.40	62.98	63.22	42.73	27.66	29.58	45.76
KG-RLVR	✓	<u>51.63</u>	67.54	67.92	<u>46.49</u>	<u>34.21</u>	<u>32.80</u>	<u>50.07</u>
†Vanilla RL	✓	65.23	86.18	73.44	71.55	50.32	52.26	66.49
†KG-SFT	✓	69.09	90.03	74.61	72.21	53.17	53.44	68.76
†KG-RLVR	✓	71.80	90.19	75.78	73.60	59.49	60.30	71.86

Table 2: Results of accuracy (%) on multi-domain QA subsets. Unless otherwise indicated (†), all other groups adopt Llama2-7B-chat as the backbone. RL is short for Vanilla RL. The best results achieved on Qwen2.5-7B-Instruct and Llama2-7B-chat are highlighted in bold and underlined.

Domain	ToG	KGR	KAPING	StructGPT	KG-SFT	RL	KG-RLVR	†RL	†KG-RLVR
Nutrition	43.55	45.16	40.32	50.84	62.35	48.72	<u>63.73</u>	63.34	73.86
Astronomy	38.71	41.94	35.48	49.50	54.84	46.71	53.94	63.16	67.11
Microeconomics	35.42	39.58	29.17	34.15	47.92	39.58	<u>49.49</u>	55.90	60.77
Formal Logic	31.12	32.65	34.42	<u>45.63</u>	39.60	34.62	42.06	56.35	63.49
Computer Security	45.00	45.00	45.00	50.00	<u>60.00</u>	46.00	58.00	71.00	80.00
Psychology	45.08	44.26	43.09	45.32	<u>52.03</u>	44.25	50.28	66.10	73.12
Professional Accounting	42.55	46.18	47.24	48.88	<u>49.62</u>	47.87	48.22	65.61	59.21
Management	57.14	52.38	57.14	68.42	75.00	67.96	<u>75.72</u>	84.47	88.35
History	52.64	51.32	49.50	47.18	<u>67.08</u>	60.23	65.68	78.71	79.54
Professional Law	34.31	34.43	32.35	36.45	43.65	40.16	<u>44.07</u>	69.06	73.01
Avg.	42.55	43.29	41.37	47.64	<u>55.21</u>	47.61	55.12	67.37	72.85

evidence units, testing the sensitivity of the reward to irrelevant or spurious KG signals.

- **NER-only supervision.** KG interaction is removed, and only NER-extracted entities are used for process supervision, testing whether surface entity signals can replace structured KG evidence.
- **No BM25 weight.** All evidence units in $\mathcal{H}$ and $\mathcal{L}$ are assigned uniform weights when computing fuzzy matching scores, testing whether BM25-based weighting provides an additional refinement beyond the partition.
- **Weighted-sum reward.** Answer-correctness gating is replaced with a weighted-sum formulation in which the KG reward contributes independently of final-answer correctness, di-

rectly testing the necessity of coupling evidence use with task success.

For the weighted-sum variant, the final reward R' is defined as

$$R' = \alpha R_{\text{format}} + (1 - \alpha)[(1 - \beta)R_{\text{acc}} + \beta R_{\text{KG}}].$$

Unlike the gated objective, this formulation allows R_{KG} to contribute even when $R_{\text{acc}} = 0$.

As shown in Table 3, all variants underperform the full KG-RLVR framework. The prompt-appended variant shows that simply exposing the model to evidence is insufficient without reward-level guidance. The no-clustering and random-evidence variants further show that noisy or irrelevant evidence can weaken training, sometimes even reducing performance below Vanilla RL. The

Table 3: Ablation results of accuracy (%) on MMedBench using Qwen2.5-7B-Instruct. For the weighted-sum reward, we test $\beta \in \{0.1, 0.2, 0.5, 0.8\}$ and report the best-performing setting, $\beta = 0.5$.

Experiment	English	Chinese	Russian	Spanish	French	Japanese	Avg.
KG-RLVR	71.80	90.19	75.78	73.60	59.49	60.30	71.86
- prompt evidence only	68.03	88.06	74.61	72.79	55.95	55.28	69.12
- w/o cluster	66.38	87.13	73.44	72.21	51.93	50.75	66.97
- random unit	65.83	87.36	73.83	71.88	49.84	52.76	66.92
- NER only	67.09	86.98	74.61	72.10	57.88	54.77	68.90
- w/o BM25	70.46	89.67	75.39	72.90	59.00	58.79	71.04
- weighted-sum reward	69.21	87.35	73.89	72.03	55.44	55.18	69.18

NER-only variant confirms that surface entity supervision cannot fully replace structured KG interaction. The no-BM25-weight variant leads to a slight performance drop, suggesting that BM25-based weighting provides an additional refinement rather than the main source of improvement. Overall, these results support the necessity of filtering, organizing, and rewarding KG-derived evidence in an RL-compatible manner.

Most importantly, the weighted-sum reward underperforms the gated objective across all six language subsets, reducing the average accuracy from 71.86 to 69.18. Because the weighted-sum formulation rewards KG evidence independently of answer correctness, it can favor trajectories that mention relevant evidence but still reach an incorrect conclusion. Its weaker performance supports coupling the KG reward with answer correctness.

3.4 Low-Relevance Evidence Study

Figure 2 examines the effect of λ , which controls the balance between high-relevance evidence coverage and low-relevance evidence discrimination in the KG reward. The results indicate that the two components should be balanced rather than optimized in isolation. Assigning all emphasis to $\mathcal{H}$ can overlook useful auxiliary clues from $\mathcal{L}$, while over-emphasizing $\mathcal{L}$ may weaken the anchoring effect of directly relevant evidence. In several language subsets, intermediate λ values achieve better performance, suggesting that KG-RLVR benefits from jointly preserving core-evidence coverage and selectively integrating less direct evidence for answer discrimination. This observation supports the motivation behind the low-relevance discrimination reward. The low-relevance set $\mathcal{L}$ should not be treated simply as noise to be removed. Instead, it can be viewed as a weakly aligned evidence pool. This is consistent with our broader view that KG-

enhanced post-training should not merely expose the model to retrieved knowledge, but should transform noisy structured evidence into rewardable reasoning preferences.

We also observe that λ has relatively limited influence on Chinese and Russian. Question length is one possible factor as shown by the average question lengths reported in Appendix B; differences in KG coverage, entity-linking quality, dataset composition, and the backbone’s multilingual capabilities may also contribute. Overall, we use $\lambda = 0.5$ as the default setting in the main experiments, which provides a balanced trade-off between core-evidence coverage and selective auxiliary-evidence use.

3.5 Evidence-Masking Analysis

The fuzzy-matching reward identifies evidence mentions in a reasoning trace, but their presence alone does not show whether they provide useful information for answering the question. We therefore conduct a post-hoc evidence-masking analysis. For each test instance, we apply the same KG retrieval and clustering procedure used during training to obtain a reference set of evidence units. We then locate the spans in the generated reasoning trace that match these units using the same fuzzy-matching function as in the reward computation and replace them with [unknown]. Conditioned on the masked trace, the model is asked to produce the final answer again. This intervention is used only for analysis and does not alter training or the standard inference protocol.

As shown in Table 4, masking the matched evidence spans reduces accuracy across all six subsets, lowering the average score from 71.86 to 64.84, even lower than Vanilla RL. This suggests that the gains from KG-RLVR are closely tied to the evidence embedded in the reasoning traces, rather than to superficial evidence mentions alone.

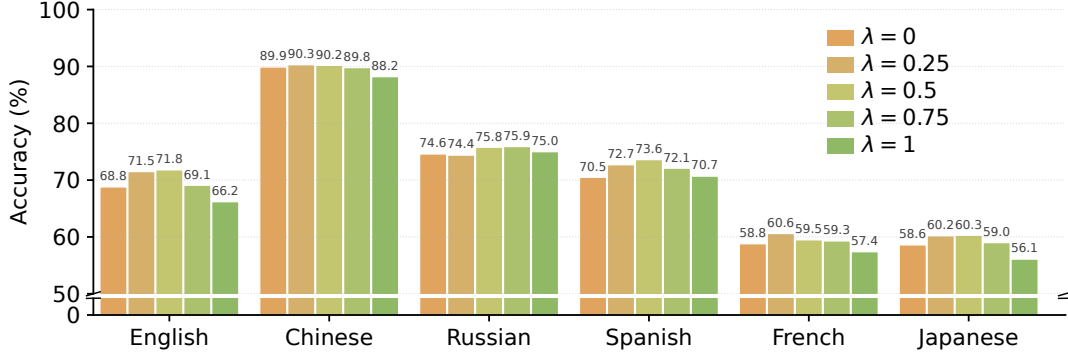


Figure 2: Effect of λ in the KG reward on MMedBench. The coefficient λ balances high-relevance evidence coverage and low-relevance evidence discrimination.

Setting	English	Chinese	Russian	Spanish	French	Japanese	Avg.
KG-RLVR	71.80	90.19	75.78	73.60	59.49	60.30	71.86
KG-RLVR with evidence masking	62.75	81.02	71.37	70.41	51.65	51.83	64.84

Table 4: Evidence-masking results of accuracy (%) on MMedBench using Qwen2.5-7B-Instruct. Matched KG-evidence spans in the generated reasoning traces are replaced with [unknown].

3.6 Case Study: Does KG-RLVR Reason Better?

KG-RLVR achieves stronger final-answer performance, but this alone does not show whether the reasoning process itself is improved. A remaining question is whether the model learns to use task-relevant knowledge more effectively, rather than merely shifting its answer distribution. To examine this issue, we evaluate whether generated reasoning traces cover evidence units that are genuinely useful for deriving the correct answer.

For 200 randomly selected MMedBench-English test instances, we obtain candidate evidence units through the same offline preprocessing procedure as in training, including NER, KG retrieval and clustering. We then use GPT-4o only as a post-hoc annotation tool to identify useful evidence units $\mathcal{E}_x^*$ for each test sample x . This annotation is used solely for analysis and is not involved in training or inference. We compute the average fuzzy matching score between generated reasoning traces and annotated useful evidence units over the test set $\mathcal{T}$:

$$\bar{m}_{\text{test}} = \frac{1}{|\mathcal{T}|} \sum_{x \in \mathcal{T}} \frac{1}{|\mathcal{E}_x^*|} \sum_{e \in \mathcal{E}_x^*} m(e, r_x).$$

This metric serves as a rule-verifiable proxy for useful evidence coverage, rather than a subjective evaluation of rationale quality. As shown in Figure 3, KG-RLVR exhibits a steady increase in this

score during training, while Vanilla RL fluctuates more strongly and may deteriorate. Since the evidence units are not provided at inference time, this suggests that KG-RLVR shapes the policy to generate reasoning traces better aligned with task-relevant KG evidence.

In addition to the rule-verifiable coverage analysis, we conduct a complementary pairwise evaluation of reasoning quality. We randomly sample 200 test instances from each language subset, except for Japanese, for which we use the entire test set of 199 instances. For each instance, GPT-4o is asked to judge which reasoning process of Vanilla RL/KG-RLVR is better, or whether the two are roughly comparable. To reduce positional bias, the two reasoning traces are presented in random order. As shown in Figure 4, KG-RLVR is preferred more often overall, suggesting that its reasoning process is also more favorably assessed by an external judge. Overall, the case study complements the accuracy results by showing that KG-RLVR improves not only the final-answer performance, but also the reasoning trace.

4 Conclusion

We presented KG-RLVR, a lightweight framework that reframes KG-enhanced LLM post-training as verifiable reward construction from traceable structured evidence. KG-RLVR abstracts retrieved KG evidence into traceable evidence units and converts them into locally computable process re-

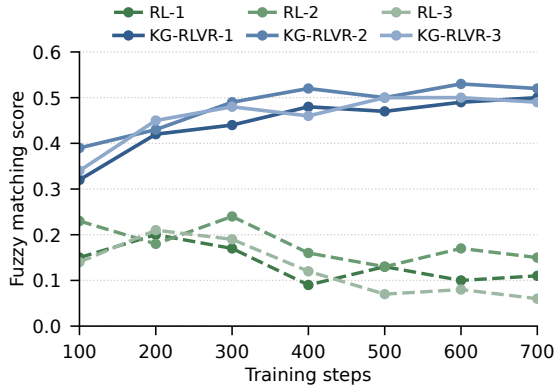


Figure 3: Fuzzy matching scores between generated reasoning traces and annotated useful evidence units on 200 randomly selected MMedBench-English test instances. Each method is evaluated with three independent random runs, shown as separate curves.

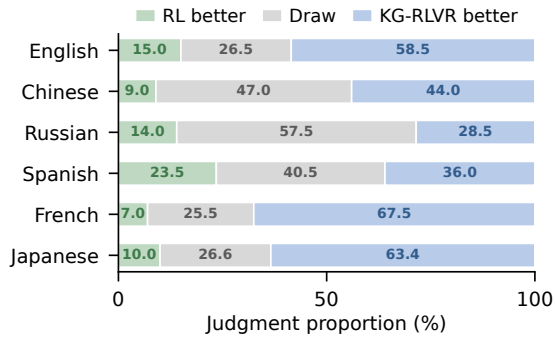


Figure 4: Pairwise GPT-4o judgments of reasoning traces generated by Vanilla RL and KG-RLVR on MMedBench. Draw indicates that the two reasoning traces are roughly comparable.

wards. Through reward-driven post-training, these rewards guide the learned policy toward reasoning traces that are more evidence-grounded, answer-supportive, and aligned with final-answer correctness. Experiments on knowledge-intensive QA tasks show that KG-RLVR consistently improves over vanilla RL, suggesting that structured and verifiable process signals provide useful guidance beyond answer-level reinforcement alone. More broadly, our results suggest that effective KG-enhanced post-training depends on transforming retrieved knowledge into optimization signals that can shape reasoning behavior. We hope this work encourages further exploration of traceable knowledge sources and knowledge-grounded reward design for reasoning-oriented LLM adaptation.

Limitations

- First, our evaluation focuses on knowledge-intensive multiple-choice QA, where final answers can be verified automatically. While KG-RLVR may extend to other tasks with verifiable outputs and explicit reasoning traces, its effectiveness on open-ended generation, multi-turn dialogue, and long-horizon reasoning remains to be established.
- Second, the current reward relies on Levenshtein-based fuzzy matching as an efficient proxy for evidence use. Although the results suggest that the matched spans are informative for answer prediction, they do not establish that the model interprets or uses the evidence correctly. Surface matching remains insensitive to paraphrases, negation, relation direction, and logically distinct concepts with similar wording. Future work could incorporate NLI- or LLM-based verification, potentially in a hybrid pipeline that applies semantic verification only to ambiguous matches to limit computational overhead.
- Finally, KG retrieval and evidence construction are performed offline. Updates to the underlying KG therefore require recomputing the affected entity links, retrieved subgraphs, and evidence units. Developing incremental preprocessing and cache-update mechanisms would make KG-RLVR better suited to evolving knowledge sources. The method also remains dependent on the coverage and accuracy of the underlying KG and entity-linking.

Ethical Statement

This work uses publicly available datasets, open-source models, and open-source software frameworks. All experiments are conducted for research purposes, and no private, personally identifiable, or sensitive user data is collected or used. Although the study is based on benchmark evaluation, models trained or evaluated in this setting may still produce incorrect, biased, or misleading outputs. Therefore, any downstream use should involve appropriate human oversight, careful domain-specific validation, and safeguards against over-reliance on automatically generated responses, especially in high-stakes scenarios.

Acknowledgments

This research was supported by grants from the National Natural Science Foundation of China (62525606), the Key Technologies R & D Program of Anhui Province (No. 202423k09020039) and the Fundamental Research Funds for the Central Universities.

References

- Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17.
- Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 78–106.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.
- Hanzhu Chen, Xu Shen, Jie Wang, Zehao Wang, Qitan Lv, Junjie He, Rong Wu, Feng Wu, and Jieping Ye. 2025. Knowledge graph finetuning enhances knowledge manipulation in large language models. In *The Thirteenth International Conference on Learning Representations*.
- Zhongwu Chen, Long Bai, Zixuan Li, Zhen Huang, Xiaolong Jin, and Yong Dou. 2024. [A new pipeline for knowledge graph reasoning enhanced by large language models without fine-tuning](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1366–1381, Miami, Florida, USA. Association for Computational Linguistics.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.
- Tiezheng Guo, Qingwen Yang, Chen Wang, Yanyi Liu, Pan Li, Jiawei Tang, Dapeng Li, and Yingyou Wen. 2024. Knowledge navigator: Leveraging large language models for enhanced reasoning over knowledge graph. *Complex & Intelligent Systems*, 10(5):7063–7076.
- Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569.
- Jiashu He, Jinxuan Fan, Bowen Jiang, Ignacio Houine, Dan Roth, and Alejandro Ribeiro. 2026. [Sake: Structured agentic knowledge extrapolation for complex llm reasoning via reinforcement learning](#). *Preprint*, arXiv:2505.15062.
- Jiashu He, Mingyu Derek Ma, Jinxuan Fan, Dan Roth, Wei Wang, and Alejandro Ribeiro. 2024. Give: Structured reasoning of large language models with knowledge graph inspired veracity extrapolation. *arXiv preprint arXiv:2410.08475*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xin Zhao, and Ji-Rong Wen. 2023. [StructGPT: A general framework for large language model to reason over structured data](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Ernests Lavrinovics, Russa Biswas, Johannes Bjerva, and Katja Hose. 2025. Knowledge graphs, large language models, and hallucinations: An nlp perspective. *Journal of Web Semantics*, 85:100844.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.
- Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2901–2908.
- Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, and 1 others. 2024a. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.
- Linhao Luo, Yuan-Fang Li, Reza Haffari, and Shirui Pan. 2024b. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*, volume 2024, pages 14400–14423.

- Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. 2025. [Large language models meet knowledge graphs for question answering: Synthesis and opportunities](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24578–24597, Suzhou, China. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Hao Peng, Yunjia Qi, Xiaozhi Wang, Bin Xu, Lei Hou, and Juanzi Li. 2025. [VerIF: Verification engineering for reinforcement learning in instruction following](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30324–30339, Suzhou, China. Association for Computational Linguistics.
- Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. [Knowledge enhanced contextual word representations](#). 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)*, pages 43–54, Hong Kong, China. Association for Computational Linguistics.
- Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.
- Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. [Sequence-to-sequence knowledge graph completion and question answering](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2814–2828, Dublin, Ireland. Association for Computational Linguistics.
- Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2025. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *International Conference on Learning Representations*, volume 2025, pages 60808–60838.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*.
- Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*, volume 2024, pages 3868–3898.
- Shiyu Tian, Yangyang Luo, Tianze Xu, Caixia Yuan, Huixing Jiang, Chen Wei, and Xiaojie Wang. 2024. [KG-adapter: Enabling knowledge graph integration in large language models through parameter-efficient fine-tuning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3813–3828, Bangkok, Thailand. Association for Computational Linguistics.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Kishore Vanapalli, Aravind Kilaru, Omair Shafiq, and Shahzad Khan. 2025. [Unifying large language models and knowledge graphs for efficient regulatory information retrieval and answer generation](#). In *Proceedings of the 1st Regulatory NLP Workshop (RegNLP 2025)*, pages 22–30, Abu Dhabi, UAE. Association for Computational Linguistics.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. [Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand. Association for Computational Linguistics.
- Shuai Wang and Yinan Yu. 2026. [Kg-reasoner: A reinforced model for end-to-end multi-hop knowledge graph reasoning](#). *Preprint*, arXiv:2604.12487.
- Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. [MindMap: Knowledge graph prompting sparks graph of thoughts in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 10370–10388, Bangkok, Thailand. Association for Computational Linguistics.

Xiaobao Wu. 2025. [A comprehensive survey on learning from rewards for large language models: Reward models and learning strategies](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17847–17875, Suzhou, China. Association for Computational Linguistics.

Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Jun Zhao, and Kang Liu. 2024. [Generate-on-graph: Treat LLM as both agent and KG for incomplete knowledge graph question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18410–18430, Miami, Florida, USA. Association for Computational Linguistics.

Lian Yan, Chen Tang, Yi Guan, Haotian Wang, Songyuan Wang, Haifeng Liu, Yang Yang, and Jingchi Jiang. 2025. [RLKGF: Reinforcement learning from knowledge graph feedback without human annotations](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6619–6633, Vienna, Austria. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. [QA-GNN: Reasoning with language models and knowledge graphs for question answering](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Kai Zhang, Qi Liu, Zhenya Huang, Mingyue Cheng, Kun Zhang, Mengdi Zhang, Wei Wu, and Enhong Chen. 2022a. Graph adaptive semantic transfer for cross-domain sentiment classification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1566–1576.

Kai Zhang, Hao Qian, Qing Cui, Qi Liu, Longfei Li, Jun Zhou, Jianhui Ma, and Enhong Chen. 2021. Multi-interactive attention network for fine-grained feature learning in ctr prediction. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 984–992.

Kai Zhang, Hefu Zhang, Qi Liu, Hongke Zhao, Hengshu Zhu, and Enhong Chen. 2019. Interactive attention transfer network for cross-domain sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5773–5780.

Kai Zhang, Kun Zhang, Mengdi Zhang, Hongke Zhao, Qi Liu, Wei Wu, and Enhong Chen. 2022b. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. *arXiv preprint arXiv:2203.16369*.

Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. 2024a. Knowgpt: Knowledge graph based prompting for large language models. *Advances in neural information processing systems*, 37:6052–6080.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022c. Greaselm: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*.

Yifei Zhang, Xintao Wang, Jiaqing Liang, Sirui Xia, Lida Chen, and Yanghua Xiao. 2024b. Chain-of-knowledge: Integrating knowledge reasoning into large language models by learning from knowledge graphs. *arXiv preprint arXiv:2407.00653*.

Ruilin Zhao, Feng Zhao, Long Wang, Xianzhi Wang, and Guandong Xu. 2024. Kg-cot: Chain-of-thought prompting of large language models over knowledge graphs for knowledge-aware question answering. In *IJCAI*, pages 6642–6650.

A Method Details

A.1 Levenshtein-Based Text Similarity

In KG-RLVR, we use a simple text similarity measure based on Levenshtein distance, i.e., edit distance. This design supports on-the-fly reward computation during RL training with low computational overhead. Given two evidence units e_1 and e_2 , we define the normalized similarity as:

$$\text{sim}(e_1, e_2) = 1 - \frac{d(e_1, e_2)}{\max(\text{len}(e_1), \text{len}(e_2))}$$

where $d(\cdot, \cdot)$ denotes the Levenshtein distance and $\text{len}(\cdot)$ denotes the string length. This similarity is used to construct the distance matrix for evidence clustering.

To match an evidence unit e against a generated reasoning trace r , we define the position-normalized similarity as:

$$\text{sim}_p(e, r) = \max_{o \in \text{windows}(r, \text{len}(e))} \left(1 - \frac{d(e, o)}{\text{len}(e)} \right)$$

where $\text{windows}(r, \text{len}(e))$ denotes all contiguous text spans in r with the same length as e . The KG matching score used in the reward is then defined as:

$$m(e, r) = \text{sim}_p^2(e, r).$$

The squared form makes the reward more sensitive to high-quality matches while suppressing weak text overlaps.

Table 5: Statistics of the multilingual MMedBench datasets. Avg. Len. denotes the average length (chars) of the question in the test set.

Dataset	Language	Source	Train	Test	Avg. Len.
MedQA	English	United States Medical Licensing Examination	10,178	1,273	745.6
MedQA	Chinese	United States Medical Licensing Examination	27,400	3,426	45.9
RuMedDaNet	Russian	Russian medical judgment question dataset	1,052	256	63.1
Head-QA	Spanish	Exams for positions in the Spanish healthcare system	2,657	2,742	144.5
FrenchMedMCQA	French	Professional exams for the French Pharmacy degree	2,171	622	99.6
IgakuQA	Japanese	Japan’s medical licensure exams	1,590	199	208.9

A.2 Evidence Clustering

We use K-medoids clustering to condense the retrieved KG-derived evidence units, as introduced in detail in Algorithm 1. Unlike K-means, K-medoids selects actual evidence units as cluster centers, allowing us to retain representative evidence units directly for reward computation.

We set the target number of clusters to:

$$N = \max(1, \lfloor 0.3|\mathcal{E}| \rfloor),$$

and set the maximum number of iterations to $T = 100$. All non-medoid evidence units are discarded after clustering and do not enter reward computation.

A.3 Reward Details

Format reward. We define the format reward using two components: a template-matching indicator and a linear length penalty. The indicator checks whether the generated response follows the required prompt template, while the length penalty is activated only when the reasoning trace exceeds a prescribed threshold. Formally, we define

$$R_{\text{format}} = \mathbb{I}(a \models \tau) + P_{\text{len}}(r),$$

where a denotes the full generated response, r denotes the reasoning trace, and $a \models \tau$ indicates that a satisfies the required prompt template τ . The length penalty is defined as

$$P_{\text{len}}(r) = \max\left(-5, -\gamma \max(0, \text{len}(r) - L_{\text{th}})\right),$$

where $\text{len}(r)$ denotes the length of the reasoning trace, L_{th} is the length threshold beyond which the penalty becomes active, and $\gamma > 0$ controls the penalty slope. We set $L_{\text{th}} = 3000$ (chars) and $\gamma = 0.002$ in practice.

BM25 weight normalization. For high-relevance evidence units, we assign a uniform

weight to emphasize full coverage of core evidence. Specifically, for each $e \in \mathcal{H}$, we set $w_e = 1$. For low-relevance evidence units, BM25 scores are used only to determine their relative importance within $\mathcal{L}$. We first compute the raw BM25 score $s(e)$ for each $e \in \mathcal{L}$ against the concatenation of the question stem and candidate options. The evidence units in $\mathcal{L}$ are then sorted in descending order of $s(e)$. Let the sorted sequence be $\{e_{(1)}, e_{(2)}, \dots, e_{(|\mathcal{L}|)}\}$, where $s(e_{(1)}) \geq s(e_{(2)}) \geq \dots \geq s(e_{(|\mathcal{L}|)})$. We assign rank-normalized weights by linearly decaying from 1 to 0.01:

$$w_{e_{(i)}} = 1 - \frac{i-1}{|\mathcal{L}|-1}(1-0.01), \quad i = 1, \dots, |\mathcal{L}|.$$

When $|\mathcal{L}| = 1$, we set $w_{e_{(1)}} = 1$. This normalization preserves the BM25-induced relevance order while avoiding excessive dependence on the absolute scale of BM25 scores.

Fallback rules. We apply simple fallback rules to ensure stable reward computation in degenerate cases. If either $\mathcal{H}$ or $\mathcal{L}$ is empty, the corresponding matching term is set to 1, which neutralizes its effect in the multiplicative KG reward. Specifically, if $\mathcal{H} = \emptyset$, we set: $\bar{m}_{\mathcal{H}} = 1$. If $\mathcal{L} = \emptyset$, we set: $\Delta \bar{m}_{\mathcal{L}} = 1$. For the low-relevance discrimination term, if $\mathcal{L}_{\text{top}}$ is empty, the weighted average over $\mathcal{L}_{\text{top}}$ is set directly to 1:

$$\frac{\sum_{e \in \mathcal{L}_{\text{top}}} w_e \cdot m(e, r)}{\sum_{e \in \mathcal{L}_{\text{top}}} w_e} = 1.$$

If $\mathcal{L}_{\text{rest}}$ is empty, the weighted average over $\mathcal{L}_{\text{rest}}$ is set directly to 0:

$$\frac{\sum_{e \in \mathcal{L}_{\text{rest}}} w_e \cdot m(e, r)}{\sum_{e \in \mathcal{L}_{\text{rest}}} w_e} = 0.$$

Algorithm 1 K-Medoids Clustering Algorithm

Require: Evidence phrase set $\mathcal{E} = \{e_1, e_2, \dots, e_M\}$, number of clusters N ($1 \leq N \leq M$), max iterations T .

Ensure: Cluster labels $L(e_i) \in \{1, \dots, N\}$ for each e_i , medoids $\{m_1, \dots, m_N\}$ where $m_j \in \mathcal{E}$.

- 1: Define clustering distance $D(x, y) = 1 - \text{sim}(x, y)$
- 2: **Initialisation:**
- 3: Select N distinct indices $i_1, \dots, i_N$ uniformly at random from $\{1, \dots, M\}$
- 4: **for** $j = 1$ to N **do**
- 5: $m_j \leftarrow e_{i_j}$
- 6: **end for**
- 7: **for** $t = 1$ to T **do**
- 8: **Assignment step:**
- 9: **for** each $e_i \in \mathcal{E}$ **do**
- 10: $L(e_i) \leftarrow \arg \min_{1 \leq j \leq N} D(e_i, m_j)$
- 11: **end for**
- 12: **Update step (PAM swap):**
- 13: $\text{changed} \leftarrow \text{False}$
- 14: **for** $j = 1$ to N **do**
- 15: $C_j \leftarrow \{e_i \in \mathcal{E} \mid L(e_i) = j\}$
- 16: **if** $C_j = \emptyset$ **then**
- 17: **continue**
- 18: **end if**
- 19: $\text{curr_med} \leftarrow m_j$
- 20: $\text{curr_cost} \leftarrow \sum_{e \in C_j} D(e, \text{curr_med})$
- 21: $\text{best_med} \leftarrow \text{curr_med}$
- 22: $\text{best_cost} \leftarrow \text{curr_cost}$
- 23: **for** each candidate $\in C_j \setminus \{\text{curr_med}\}$ **do**
- 24: $\text{cand_cost} \leftarrow \sum_{e \in C_j} D(e, \text{candidate})$
- 25: **if** $\text{cand_cost} < \text{best_cost}$ **then**
- 26: $\text{best_cost} \leftarrow \text{cand_cost}$
- 27: $\text{best_med} \leftarrow \text{candidate}$
- 28: $\text{changed} \leftarrow \text{True}$
- 29: **end if**
- 30: **end for**
- 31: $m_j \leftarrow \text{best_med}$
- 32: **end for**
- 33: **if** $\neg \text{changed}$ **then**
- 34: **break**
- 35: **end if**
- 36: **end for**
- 37: **return** $\{L(e_i)\}_{i=1}^M$ and $\{m_1, \dots, m_N\}$

Table 6: Subset sizes of different domains.

Subset	Size
Nutrition	306
Astronomy	152
Microeconomics	390
Formal Logic	126
Computer Security	100
Psychology	545
Professional Accounting	282
Management	103
History	606
Professional Law	1,534
Auxiliary Train	99,842

Table 7: Key Hyperparameter settings.

Hyperparameter	Value
Learning rate	1×10^{-6}
Learning rate (Vanilla RL)	3×10^{-7}
Batch size	64
Total epochs	5
α	0.2
β	0.8
λ	0.5
r_k	0.15
Rollout	5
Entropy coef.	0.001
KL loss coef.	0.001
Max prompt len.	2048
Max response len.	2048
Max model len.	4096
GPU memory utilization	0.5

B Experiment Details

B.1 Datasets

MMedBench is a medical-domain multiple-choice QA benchmark aggregated from multiple source datasets. Each instance contains a question, candidate options, the correct answer, and a pseudo-rationale generated by GPT-4. In our experiments, all pseudo-rationales are discarded and not used for training. In addition, the data leakage of this benchmark is very low, as reported by [Chen et al. \(2025\)](#). The dataset sizes are shown in Table 5.

We use the MMLU subset as the multi-domain question answering benchmark. Following the baseline setting, we train the model on the auxiliary training set and evaluate it on selected test

subsets. The sizes are shown in Table 6.

B.2 Training Implementation

Our implementation of KG-RLVR is built on the open-source RL framework VeRL-0.4.1 ([Sheng et al., 2024](#)) and vLLM-0.8.4 ([Kwon et al., 2023](#)). Since VeRL supports user-defined reward functions, KG-RLVR only requires implementing the proposed KG-guided reward within the reward computation module, making it straightforward to integrate into the existing RL training pipeline. We provide a repository that includes the core training pipeline, the KG-guided reward computation module, RL training scripts, and configuration files. The key hyperparameters are listed in Table 7. All the RL experiments were executed on 4x Tesla

A800-80GB GPUs. For our RL-based experiments, accuracy results are reported as the median across three independent random runs.

B.3 Prompt Template

» NER Prompt

Please extract all medically relevant knowledge items and concepts from the following text. Return only the extracted items, without type labels or additional explanations. Place each item on a separate line, use no more than 10 words per item, and avoid symbols. End the output with ### after the final item. Extract at least three items.

Text: {text}

The medical items are as follows:

» Data4training Prompt

{question}

Options:

{options}

Analyze the question step by step and then provide the final answer option. Format:

```
<think>
reasoning here
</think>
<answer>
answer option
</answer>
```

» Useful Evidence Selection Prompt

Given a medical multiple-choice question, its candidate options, the correct answer, and a list of candidate KG-derived evidence units, identify which evidence units are genuinely useful for deriving the correct answer.

Question:

{question}

Options:

{options}

Correct Answer:

{answer}

Candidate Evidence Units:

{evidence_units}

Select only the evidence units that are directly helpful for solving the question. Do not select evidence that is merely related to the topic but not useful for deriving the answer. Return only the selected evidence units, one per line, without explanations. End the output with ### after the final selected unit.

» Pairwise Reasoning Evaluation Prompt

Given a medical multiple-choice question, its candidate options, and two reasoning traces, compare which reasoning trace better supports solving the question.

Question:

{question}

Options:

{options}

Reasoning Trace A:

{reasoning_a}

Reasoning Trace B:

{reasoning_b}

Evaluate the two reasoning traces based on their relevance to the question, correctness of medical knowledge, logical coherence, and usefulness for deriving the final answer. If one reasoning trace is clearly better, output A or B. If they are roughly comparable, output Draw. Return only one label from {A, B, Draw}, without additional explanations.

For multi-domain data, we use the same template with medical-domain terms replaced by general-domain terms.

C More Related Work

Integrating KG with pretrained language models has long been an active area of research (Peters et al., 2019; Zhang et al., 2019; Liu et al., 2020; Yasunaga et al., 2021; Zhang et al., 2021, 2022b; Saxena et al., 2022). Building on this long-standing line of work, recent LLM-KG research has increasingly focused on how to translate structured facts, relational constraints, and interpretable graph paths into forms that LLMs can effectively use to improve their predictions and reasoning. In prompt- and retrieval-augmented paradigms, KAPING retrieves relevant facts from KGs based on semantic similarity and prepends them to the prompt, enabling training-free KGQA (Baek et al., 2023). StructGPT abstracts question answering over structured data into an Iterative Reading-then-Reasoning framework, emphasizing tool-like access to structured data (Jiang et al., 2023). KnowGPT and MindMap further attempt to improve the search space, prompt construction, and transparency of KG-augmented reasoning (Zhang et al., 2024a; Wen et al., 2024). ToG treats the LLM as an agent that performs beam search over the KG, thereby directly embedding graph traversal into the reasoning process (Sun et al., 2024). RoG emphasizes a planning-retrieval-reasoning paradigm, regarding relation paths as faithful plans (Luo et al.,

2024b). KnowledgeNavigator organizes reasoning through a pipeline of question-constraint augmentation, iterative graph retrieval, and structured prompt construction (Guo et al., 2024). KG-CoT first employs a lightweight graph reasoner to produce high-confidence reasoning paths, which are then passed to a large language model for answer generation (Zhao et al., 2024).

In terms of training-time integration, GreaseLM represents an early line of work in the LM era that performs multi-layer interaction and fusion between LM representations and graph neural network representations (Zhang et al., 2022c). KG-Adapter further extends KG injection to decoder-only LLMs (Tian et al., 2024). Chain-of-Knowledge constructs the KnowReason dataset via KG rule mining and combines it with a trial-and-error mechanism to train models' knowledge reasoning capabilities (Zhang et al., 2024b). KG-SFT advances this line of research toward domain adaptation, arguing that high-quality KG-derived explanations can improve a model's ability to understand and manipulate the associations and logic underlying QA instances (Chen et al., 2025).

Notably, recent work has also made substantial progress in KG-RL. RLKGF proposes that KGs can provide external feedback for RL when human annotations are unavailable (Yan et al., 2025). SAKE uses tool-augmented RL to perform knowledge extrapolation over KGs (He et al., 2026). KG-Reasoner directly internalizes KG traversal into the "thinking" stage of reasoning LLMs, enabling models to learn how to actively retrieve knowledge during reasoning (Wang and Yu, 2026). Taken together, these studies indicate that KGs are no longer merely external knowledge stores for LLMs, but can serve as actionable supervision for shaping how models retrieve, select, and use knowledge during reasoning—particularly when explicit process annotations are costly or unavailable.

D Case Presentations

In this section, we present three model-output cases drawn from the MMedBench-English test set. The first two show that KG-RLVR produces substantive reasoning that integrates relevant medical concepts, suggesting that its longer traces reflect meaningful coverage rather than mere verbosity. The third reveals a limitation: the model overweights the patient's haloperidol use while underweighting the abrupt intraoperative onset and physiolog-

ical evidence supporting malignant hyperthermia. This failure further highlights that surface-level evidence matching cannot fully capture the relative importance and causal roles of different clinical cues. The cases are listed as follows:

Case 1 (Correct).

Question. A 42-year-old male presents to his primary care physician complaining of fatigue. He has not seen a doctor since age 22. He reports that over the past three months, he has felt tired and weak despite no changes in diet or exercise. He is otherwise healthy and takes no medications. Family history is notable for colorectal cancer in his father and paternal uncle, ovarian cancer in his paternal grandmother, and pancreatic cancer in his paternal uncle. Physical examination is notable for conjunctival pallor. A complete blood count reveals a hemoglobin of 9.1 g/dL and hematocrit of 31%. A stool sample is hemoccult positive and a colonoscopy reveals a fungating hemorrhagic mass in the ascending colon. Which of the following processes is most likely impaired in this patient?

Options. (A) Base excision repair; (B) Nucleotide excision repair; (C) Mismatch repair; (D) Non-homologous end joining.

Model output. <think> Identifying hereditary cancer syndromes often depends on recognizing characteristic tumor spectra and linking them to specific DNA maintenance pathways.

A. Base excision repair. Repairs small base lesions from oxidation or alkylation, initiated by DNA glycosylases. Biallelic MUTYH mutations cause MAP (MUTYH-associated polyposis), but that phenotype lacks the ovarian/pancreatic extras. The BER family also includes enzymes like OGG1 and NTHL1 that handle oxidative guanine damage, yet none of these tie into Lynch syndrome.

B. Nucleotide excision repair. Removes bulky UV-induced adducts (e.g., pyrimidine dimers). Deficient NER gives rise to xeroderma pigmentosum or Cockayne syndrome, not hereditary colon cancer.

C. Mismatch repair. Corrects base-base mismatches and insertion/deletion loops during replication. Germline losses in MLH1 or MSH2 drive microsatellite instability (MSI-H)—the molecular fingerprint of Lynch syndrome, perfectly matching this right-sided tumor and family trio (colon, ovary, pancreas).

D. Non-homologous end joining. Repairs double-strand breaks via direct ligation, requiring the Ku70/Ku80 heterodimer and DNA-PKcs. Defects in ATM cause ataxia-telangiectasia, while LIG4 mutations produce LIG4 syndrome with radiosensitivity—but neither presents with this adenocarcinomatous spectrum.

In summary, the correct option is C. Mismatch repair. </think> <answer>C </answer>

Correct answer: C. Mismatch repair.

Case 2 (Correct).

Question. A 29-year-old woman presents to her primary care doctor for a prenatal visit. She has just learned that she is pregnant and estimates she is approximately 6 weeks pregnant given the timing of her last menstrual cycle. She has a history of recurrent fetal loss, with 4 miscarriages between gestational ages of 8 and 15 weeks. She has a history of systemic lupus erythematosus and has had several DVTs in the past and a clot in her lung once. Her vitals today are unremarkable. She is asking if there are any treatments she should receive to prevent fetal loss. Which of the following is the most appropriate management of this patient?

Options. (A) Aspirin and heparin; (B) Aspirin and warfarin; (C) No treatment indicated; (D) Rivaroxaban.

Model output. <think> High-risk pregnancy management requires integrating thromboembolic risk stratification with the fetoplacental safety profiles of available anticoagulants.

A. Aspirin and heparin. Low-dose aspirin plus LMWH (enoxaparin) is the obstetric APS mainstay. LMWH does not cross the placenta and carries no teratogenic signal, effectively reducing placental infarction.

B. Aspirin and warfarin. Warfarin crosses the placenta freely and during weeks 6–12 causes fetal warfarin syndrome (chondrodysplasia punctata, nasal hypoplasia). Although warfarin dosing is influenced by CYP2C9 and VKORC1 polymorphisms, those genetic nuances do not override its absolute contraindication in this first-trimester scenario.

C. No treatment indicated. With her history (≥ 3 losses and prior PE/DVT), this is high-risk APS; expectant management would yield a fetal loss rate exceeding 50%. Intervention is mandatory.

D. Rivaroxaban. A direct factor Xa inhibitor (DOAC). While its levels can be monitored via anti-Xa chromogenic assays, pregnancy safety data remain insufficient; moreover, DOACs (including apixaban and edoxaban) have shown inferiority to heparin in high-risk APS, especially in triple-positive patients.

Thus, the best answer is A. Aspirin and heparin.

Correct answer: A. Aspirin and heparin.

Case 3 (Failure).

Question. A 15-year-old boy is brought to the emergency department because of severe abdominal pain and vomiting for 8 hours. He has had around 3–4 episodes of greenish colored vomit during this period. He has major depressive disorder with psychosis. His mother has Graves' disease. Current medications include sertraline and haloperidol. He appears uncomfortable. His temperature is 37.3°C (99.1°F), pulse is 87/min, and blood pressure is 118/72 mm Hg. He is diagnosed with acute appendicitis and is taken to the operating room for an emergency laparoscopic appendectomy. The appendix is gangrenous and perforated with purulent fluid in the pelvis. Just after the appendiceal base is ligated, the

patient shows muscle rigidity and profuse diaphoresis. His temperature is 39.1°C (102.4°F), pulse is 130/min, and blood pressure is 146/70 mm Hg. The pupils are equal and reactive to light. The end-tidal CO₂ is 85 mm Hg. Which of the following is the most appropriate treatment for this patient's condition?

Options. (A) Dantrolene; (B) Bromocriptine; (C) Cyproheptadine; (D) Surgical embolectomy.

Model output. <think> Perioperative hyperthermic crises require consideration of both medication-related syndromes and anesthetic triggers, with particular attention to acute physiological deterioration patterns.

A. Dantrolene. Dantrolene is specific to malignant hyperthermia (MH), triggered by volatile anesthetics or succinylcholine—yet no such agents are explicitly listed here. The patient's chronic haloperidol therapy provides a far more direct etiological link to hyperthermic syndromes. Moreover, MH typically presents with masseter spasm immediately upon anesthetic induction, whereas this patient's rigidity emerges after ligating the appendix, which could simply reflect surgical stress unmasking an underlying NMS rather than a primary anesthetic reaction. Given the dominant antipsychotic exposure, dantrolene is not indicated.

B. Bromocriptine. Bromocriptine is a dopamine agonist that reverses D₂ receptor blockade—the central pathophysiological defect in neuroleptic malignant syndrome (NMS). The patient's haloperidol use provides an unambiguous trigger. NMS classically presents with “lead-pipe” rigidity, hyperthermia, autonomic lability (hypertension, tachycardia, diaphoresis), and fluctuating consciousness—all features present here. The ETCO₂ of 85 mm Hg is readily explained by sustained muscle contraction generating excess CO₂, combined with restrictive ventilatory impairment from chest-wall rigidity; this is not exclusive to MH. His mother's Graves' disease is a red herring, perhaps suggesting familial metabolic susceptibility, but does not detract from the NMS diagnosis. Bromocriptine, initiated promptly, would restore central dopaminergic tone and gradually abate the syndrome.

C. Cyproheptadine. A 5-HT_{2A} antagonist for serotonin syndrome, which could be suspected due to sertraline (SSRI) use. However, serotonin syndrome manifests with clonus, hyperreflexia, myoclonus, and diarrhea—none observed. His pupils are equal and reactive, arguing against severe serotonergic overactivity (which typically causes mydriasis). Thus, serotonin syndrome is unlikely, and cyproheptadine is inappropriate.

D. Surgical embolectomy. PE during pelvic surgery is a theoretical concern, but hypertension and elevated ETCO₂ directly contradict PE physiology (hypotension, decreased ETCO₂). Moreover, PE never causes generalized rigidity. This option is clearly irrelevant.

Therefore, the answer to this question is B. Bromocriptine.

Correct answer: A. Dantrolene.