SJCS Scientific Journal of Computer Science

futuristech PT. Teknologi Futuristik Indonesia

Date of publication March 2, 2026, date of current version March 2, 2026.
Digital Object Identifier 10.64539/sjcs.v2i1.2026.380

Article

A Self-Reflection Mechanism for Reducing Hallucination in Vietnamese Legal Question Answering Systems

Thi Vuong Pham[1,*], Nguyet Minh Phan[2], Bao Quynh Cao[2], Cong Phuc Truong[2], Thanh Duy Nguyen[2], Minh Vy Tien[2], Ngoc Son Nguyen[2]

[1] Office of Personnel Organization, Saigon University, Ho Chi Minh City 700000, Vietnam; vuong.pham@sgu.edu.vn
[2] Faculty of Information Technology, Saigon University, Ho Chi Minh City 700000, Vietnam; minhpn@sgu.edu.vn, baoquynhpy123@gmail.com, truongphuc056@gmail.com, contact.duynguyen@gmail.com, tienminhvydev@gmail.com, sonnguyen.20050319@gmail.com
* Correspondence

The authors received no financial support for the research, authorship, and/or publication of this article.

Abstract: Legal question answering is essential for compliance, dispute resolution, and everyday HR decision-making, yet large language models may produce persuasive but incorrect legal statements when supporting evidence is incomplete. While Retrieval-Augmented Generation and graph-based retrieval can ground responses in statutes and structured relations, Vietnamese legal QA often lacks an explicit, automated quality-control step that scores an answer, decides whether it should be refined, and checks that citations are actually supported. In this paper, we propose a self-reflection mechanism that adds an iterative generate–evaluate–refine loop to a Graph-RAG pipeline for Vietnamese labor-law questions. Each draft is evaluated with a hybrid score that combines how closely the answer matches retrieved legal context with a model-derived confidence estimate, and the system iterates until it reaches a quality threshold or a stopping limit. On a Vietnamese Labor Law benchmark, the approach improves accuracy from 81.5% to 86.7% and reduces hallucination from 18.7% to 9.3%, with only a modest increase in end-to-end latency in typical use. We also examine component contributions and remaining failure cases, finding that pairing contextual alignment with confidence produces more stable answers than relying on a single signal. These results indicate that self-reflection can serve as a lightweight, deployment-friendly safety layer for high-stakes legal QA without requiring additional labeled data or model fine-tuning, and it can be adapted to other Vietnamese legal domains that demand transparent, article- and clause-level evidence.

Keywords: Hallucination reduction; Knowledge graph; Large language models; Legal question answering; Retrieval-augmented generation; Self-reflection; Vietnamese natural language processing.

1. Introduction

The integration of Large Language Models (LLMs) [1] with Retrieval-Augmented Generation (RAG) has significantly advanced legal question answering systems by grounding generated answers in retrieved evidence and enabling traceable citations [2]–[4]. Graph-RAG further strengthens this paradigm by leveraging knowledge graph structures to encode legal document hierarchies and entity–relation semantics, supporting multi-hop retrieval and better contextual coverage [5], [6].

Despite these advances, Vietnamese legal QA remains difficult to deploy safely in practice. Hallucination—fluent statements that are unsupported by the retrieved context or the underlying law—can mislead users and introduce legal risk [7], [8]. Moreover, response quality can be inconsistent across runs, especially for multi-article questions and ambiguous queries, because the system lacks an explicit mechanism to assess whether the current answer is sufficiently grounded before returning it to the user [9].

Self-reflection is an emerging technique where an LLM critiques and refines its own outputs through iterative feedback, improving reliability without model fine-tuning or additional labeled data [10], [11]. However, existing self-reflection work is rarely adapted to Vietnamese legal QA with (i) graph-based retrieval, (ii) an explicit,

reproducible confidence signal, and (iii) citation verification aligned with Vietnam's legal drafting hierarchy [12].

In this paper, we introduce a self-reflection mechanism for Graph-RAG in Vietnamese Labor Law QA (Labor Code 2019 [13]). We propose a hybrid scoring mechanism that combines semantic similarity with a normalized confidence estimate and triggers iterative refinement only when needed. This design explicitly targets hallucination reduction and stabilizes response quality while controlling latency through an iteration cap.

The contribution of this research is as follows:

1) A self-reflection mechanism for Graph-RAG that reduces hallucination via an iterative generate–evaluate–refine loop with bounded iterations, designed for Vietnamese legal QA;
2) A reproducible hybrid scoring function that combines SBERT-based semantic similarity with an explicit confidence estimate (critic logit + sigmoid normalization) for quality gating;
3) Threshold sensitivity analysis justifying $\theta = 0.75$ and comprehensive ablations demonstrating the impact of similarity, confidence, and weighting on performance;
4) Experimental validation on Vietnamese Labor Law documents showing improved accuracy (81.5% → 86.7%) and reduced hallucination (18.7% → 9.3%) with moderate latency overhead.

The remainder of this paper is organized as follows. Section 2 and Section 3 presents the related work and methodology including system architecture, and the self-reflection mechanism. Section 4 presents experimental results and discussion. Section 5 summarizes limitations and future work. Section 6 concludes the paper.

2. Related Work

2.1. Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) was introduced by Lewis et al. [2] as a paradigm for combining retrieval systems with generative language models. The key insight is that LLMs can be augmented with external knowledge at inference time, allowing them to generate responses grounded in retrieved documents rather than relying solely on parametric knowledge [3]. This approach addresses the knowledge cutoff problem and enables the system to incorporate up-to-date information without retraining [14].

Karpukhin et al. [4] developed Dense Passage Retrieval (DPR), enabling efficient semantic search through dense vector representations. Unlike traditional keyword-based methods like BM25 [15], DPR uses neural encoders to map queries and documents into a shared embedding space [4]. Graph-RAG extends the RAG paradigm by incorporating knowledge graph structures [5], [6]. Yang et al. [5] proposed structure-oriented RAG that leverages knowledge graphs for co-learning with LLMs, demonstrating that explicit entity-relationship structures can enhance retrieval precision [16].

2.2. Self-Reflection in Large Language Models

Self-reflection mechanisms enable LLMs to evaluate and improve their outputs through introspection [10], [11], [17]. Shinn et al. [10] introduced Reflexion, demonstrating that LLMs can learn from their mistakes through verbal reinforcement without parameter updates. The key innovation is using natural language feedback to guide iterative improvement [18].

Madaan et al. [11] proposed Self-Refine, an iterative framework for output improvement without supervised training. Their approach uses the same LLM for both generation and critique, creating a feedback loop that progressively improves output quality [19]. Recent work has explored confidence calibration [20], factual verification [21], and multi-agent debate [22]. Pan et al. [23] surveyed fact-checking approaches for LLMs, while Huang et al. [24] reviewed reasoning capabilities. Our work adapts these principles for the legal domain with domain-specific scoring mechanisms.

2.3. Vietnamese Legal Natural Language Processing

Vietnamese legal NLP presents unique challenges due to the language's morphological complexity and specialized legal terminology [25]. The Vietnamese legal system follows a hierarchical document structure defined by Circular 25/2011/TT-BTP [12]. As illustrated in Figure 1, documents are organized hierarchically from Document (Văn bản) to Chapter (Chương), Section (Mục), Article (Điều), Clause (Khoản), and Point (Điểm).

Pham et al. [26] developed the Legal-Onto model for Vietnamese traffic law documents. Nguyen et al. [27] built intelligent search systems for Vietnamese labor law using knowledge graph techniques. Dang et al. [28] integrated knowledge graphs with LLMs and RAG for legal query systems. For Vietnamese NLP infrastructure, PhoBERT [29] provides pre-trained language models specifically designed for Vietnamese. VnCoreNLP [30] offers a comprehensive toolkit for Vietnamese text processing. Recent advances include ViHealthBERT [31] for healthcare and multilingual approaches [32].

2.4. Hallucination in Large Language Models

Hallucination refers to the generation of content that is not supported by the input context or factual knowledge [7], [8]. In legal applications, hallucination is particularly problematic because incorrect legal information can lead to serious consequences [33]. Ji et al. [7] provided a comprehensive survey of hallucination in natural language generation. Zhang et al. [34] analyzed hallucination patterns in retrieval-augmented systems,

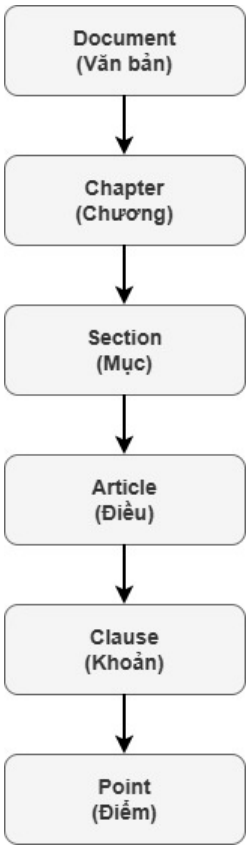

Figure 1. Vietnamese legal document hierarchy (Document → Chapter → Section → Article → Clause → Point) based on Circular 25/2011/TT-BTP.

while Manakul et al. [35] proposed detection methods. Our self-reflection mechanism contributes to this line of work by enabling the system to detect and correct hallucinated content through iterative verification.

3. Method

3.1. System Architecture

The self-reflection mechanism operates as an iterative loop that evaluates and refines LLM responses until quality criteria are met. Given a user query q and retrieved context C from the Graph-RAG system, the process proceeds through four stages: (1) Initial Generation - the LLM generates response r_0 based on query and context; (2) Quality Assessment - the system computes hybrid quality score S(r); (3) Threshold Check - if S(r) ≥ θ, accept response; otherwise proceed to refinement; (4) Iterative Refinement - identify weaknesses and regenerate with enhanced context, for maximum n iterations. Figure 2 summarizes the overall self-reflection workflow.

3.2. Hybrid Scoring Formula

We propose a hybrid scoring formula that combines two complementary quality indicators: semantic similarity and model confidence. Each metric captures different aspects of response quality, and their combination provides more robust evaluation than either metric alone [36].

3.2.1. Semantic Similarity Score

The similarity score measures semantic alignment between the generated response r and the retrieved context C using Sentence-BERT (SBERT) embeddings [37] and cosine similarity:

$$sim(r, C) = \cos(v_r, v^c) = \frac{v_r \cdot v^c}{||v_r|| \times ||v^c||} \quad (1)$$

where v_r and v^c are the SBERT embedding vectors for the response and context respectively. This metric ensures that the generated response is semantically grounded in the retrieved legal content, helping detect responses that deviate from the source material [38].

3.2.2. Confidence Score

The confidence score uses the sigmoid function to normalize raw confidence values to the bounded range (0, 1) [39]:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

The raw confidence value x is produced by an auxiliary critic step. After generating an answer r from the LLM, we prompt the same model (temperature = 0, constrained output format) to output a single scalar confidence logit x ∈ [−5, 5] indicating how likely the answer is fully supported by the retrieved context C and includes valid legal citations. We then normalize it using the sigmoid function σ(x) to obtain conf(r) ∈ (0, 1). This design avoids dependence on model-internal token logits while remaining reproducible across closed-source LLM APIs.

3.2.3. Combined Hybrid Score

The final hybrid score combines similarity and confidence with weighted parameters:

$$S(r) = \alpha \times sim(r, C) + \beta \times conf(r) \quad (3)$$

where α = 0.4 and β = 0.6 are empirically determined weights. The higher weight on confidence reflects the importance of model certainty in legal applications [40]. These weights were optimized through grid search on a validation set, testing combinations in increments of 0.1.

3.3. Quality Threshold and Iteration Control

The quality threshold θ controls whether a response is acceptable for delivery. We select θ through a sensitivity sweep (θ ∈ {0.60, 0.65, 0.70, 0.75, 0.80, 0.85}) and compute the F1-score of the accept/reject decision (accepting correct answers vs rejecting/iterating on incorrect ones). As shown in Figure 3, θ = 0.75 yields the highest F1-score, balancing precision and recall while avoiding excessive iterations.

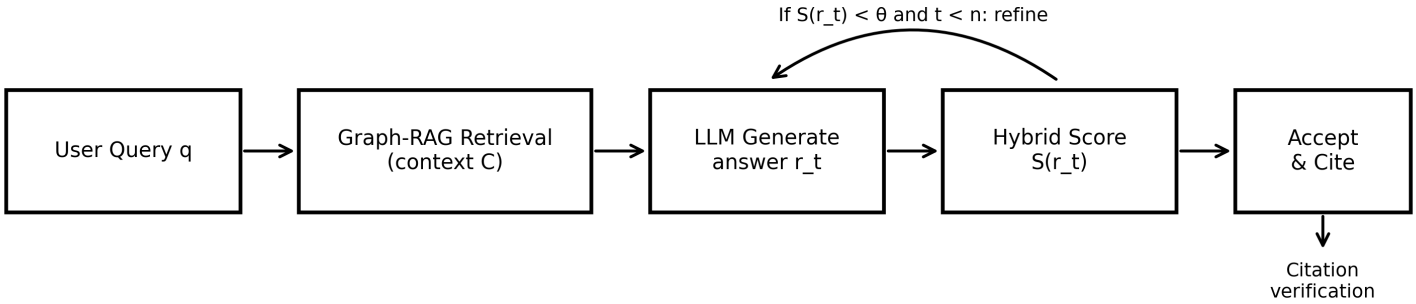

Figure 2. Self-Reflection Mechanism: The iterative loop processes queries through initial generation, quality assessment, threshold checking, and refinement stages.

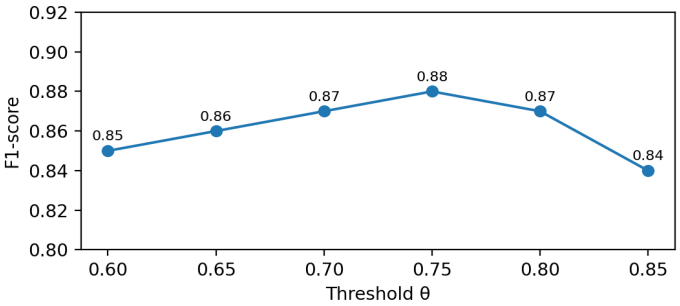

Figure 3. Threshold sensitivity analysis for selecting θ: F1-score of the accept/reject decision versus threshold θ. The best performance is achieved at θ = 0.75.

Table 1. Dataset Statistics for Vietnamese Labor Law.

Metric	Value
Number of Documents	18
Legal Concepts	489
Relationships	1,876
Knowledge Graph Triplets	3,245
Test Queries	200

Maximum iterations n = 3 prevents infinite loops while allowing sufficient refinement opportunities. Our analysis shows that most quality improvements occur in the first two iterations, with diminishing returns thereafter [41]. The iteration limit also ensures bounded response latency for real-world deployment.

3.4. Refinement Process

When the quality score falls below the threshold, the system enters a refinement phase consisting of three subprocesses: (1) Weakness Identification - analyzing the response for factual inconsistencies, missing information, incomplete citations, and logical gaps; (2) Context Enhancement - retrieving additional context from the knowledge graph to address identified gaps; (3) Regeneration - generating a new response with an enhanced prompt that explicitly addresses the identified issues [11].

3.5. Citation Verification

After a response passes the quality threshold, citation verification ensures that all referenced legal provisions are accurate. The system extracts cited articles (Điều), clauses (Khoản), and points (Điểm) from the Labor Code 2019 [13] and cross-references them against the knowledge graph [26]. This is particularly important for Vietnamese legal documents due to their hierarchical structure following Circular 25/2011/TT-BTP [12].

4. Results and Discussion

4.1. Experimental Setup

4.1.1. Dataset

We evaluate the self-reflection mechanism on Vietnamese Labor Law documents, specifically the Labor Code 2019 (Bộ luật Lao động số 45/2019/QH14) [13], one of the most frequently consulted legal areas in Vietnam. The dataset comprises 18 legal documents including the main Labor Code and related decrees (Nghị định 145/2020/NĐ-CP, Nghị định 152/2020/NĐ-CP) and circulars. The knowledge graph contains 489 legal concepts and 1,876 relationships, covering topics such as labor contracts (hợp đồng lao động), worker rights (quyền người lao động), wages (tiền lương), working hours (thời giờ làm việc), and social insurance (bảo hiểm xã hội). Dataset statistics are summarized in Table 1.

4.1.2. Baselines and Metrics

We compare against four baselines: (1) Keyword Search using the BM25 ranking function [15], grounded in classical information retrieval principles [42], [43]; (2) Pure LLM (GPT-4) without retrieval augmentation; (3) Dense RAG using SBERT embeddings [37]; (4) Graph-RAG without self-reflection [5]. Evaluation metrics include Accuracy, Hallucination Rate, Average Iterations, and Latency (seconds). For threshold selection, we also report precision, recall, and F1-score of the accept/reject decision.

Accuracy:

$$Acc = \frac{1}{N}\sum_{i=1}^{N} 1[answer_i \text{ is correct}] \quad (4)$$

Hallucination rate:

$$Hall = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[answer_i \text{ contains unsupported claims}] \quad (5)$$

Table 2. Main Experimental Results Comparing Methods (Avg.Attempts reports the mean number of generation attempts including the initial answer; "–" indicates non-iterative baselines.).

Method	Accuracy	Halluc.%	Avg. Attempts	Latency
Keyword Search	62.3%	38.2%	-	0.8s
Pure LLM	71.5%	28.5%	-	2.1s
Dense RAG	78.9%	24.5%	-	2.1s
Graph-RAG (base)	81.5%	18.7%	-	3.2s
Proposed Method	86.7%	9.3%	1.8	4.5s

Table 3. Ablation Study: Contribution of Scoring Components.

Configuration	Accuracy	Δ vs Full
Graph-RAG (base) without self-reflection	81.5%	-5.2%
Similarity only (α=1.0, β=0.0)	81.2%	-5.5%
Confidence only (α=0.0, β=1.0)	79.8%	-6.9%
Equal weights (α=0.5, β=0.5)	84.5%	-2.2%
Full (α=0.4, β=0.6)	86.7%	-

Precision, Recall, and F1-score:

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (8)$$

4.2. Main Results

As shown in Table 2, the proposed method achieves 86.7% accuracy, representing a 5.2 percentage point improvement over the base Graph-RAG system (81.5%). More importantly, the hallucination rate drops from 18.7% to 9.3%, a reduction of 9.4 percentage points (-50.3% relative reduction). This substantial decrease is critical for legal applications where factual accuracy is essential [33].

The average iteration count of 1.8 indicates that most queries require 1-2 refinement cycles. Approximately 45% of queries pass the quality threshold on the first attempt, 35% require one iteration, and 20% require two or three iterations. Latency increases from 3.2s to 4.5s (+40.6%), which we consider an acceptable trade-off given the significant quality improvement.

4.3. Ablation Study

To understand the contribution of each component, we conduct ablation experiments removing or isolating individual scoring metrics.

The ablation study in Table 3 reveals several important findings. First, both similarity and confidence contribute to overall performance: using only similarity (α=1.0, β=0.0) or only confidence (α=0.0, β=1.0) reduces accuracy by 5.5 and 6.9 percentage points respectively compared to the full configuration. Second, the hybrid approach with optimized weights (α=0.4, β=0.6) outperforms equal weighting by 2.2 percentage points, suggesting that confidence is a stronger indicator of answer correctness in our legal QA setting. These results support the design choice of emphasizing confidence during the acceptance decision.

4.4. Iteration Analysis

We analyze how performance evolves across iterations. Starting from the base Graph-RAG accuracy of 81.5% (iteration 0), accuracy improves to 84.1% after the first iteration (+2.6%), 85.6% after the second (+1.5%), and 86.7% after the third (+1.1%). The diminishing returns pattern suggests that our choice of n = 3 maximum iterations is appropriate; additional iterations would provide marginal improvement while increasing latency and the risk of over-editing.

With threshold θ=0.75, the iteration distribution is: 45% terminate after initial response, 35% after one iteration, 15% after two iterations, and only 5% require the full three iterations. This distribution demonstrates that the self-reflection mechanism is efficient, applying intensive refinement only to difficult queries.

We further discuss limitations and future work in Section 4, including risks of over-iteration and dependency on knowledge graph coverage.

4.5. Error Analysis

We analyze the remaining errors (13.3% of queries) to identify limitations. Errors fall into three categories: (1) Knowledge Gap (42% of errors) - queries requiring information not present in the knowledge graph, such as recent amendments to the Labor Code; (2) Complex Reasoning (35% of errors) - multi-hop queries requiring rea-

soning across multiple articles of the Labor Code [13]; (3) Ambiguous Queries (23% of errors) - queries with multiple valid interpretations.

4.6. Discussion

Our results have several implications for LLM-based legal question answering. First, self-reflection is effective without model retraining, making it practical for deployment [41]. Second, hybrid scoring outperforms single-metric approaches, suggesting response quality is multidimensional [36]. Third, the 40% latency increase is acceptable for legal applications where accuracy is paramount.

5. Limitations and Future Work

Limitations: First, our mechanism depends on the coverage and correctness of the underlying knowledge graph; missing or outdated provisions can still lead to incomplete answers. Second, iterative self-reflection increases inference cost and latency, which may be unsuitable for strict real-time settings. Third, the hybrid score parameters (α, β, θ) may require light re-tuning when transferring to a new legal domain or a different LLM.

Risk of over-iteration/model drift: Excessive reflection cycles can over-edit an initially correct answer, introduce stylistic drift, or amplify spurious constraints. We mitigate this by (i) enforcing a small maximum iteration cap ($n = 3$), (ii) using a monotonic acceptance rule (stop as soon as $S(r) \geq \theta$), and (iii) performing citation verification after acceptance to prevent drift away from grounded legal clauses.

Future work: We plan to extend the approach to additional Vietnamese legal domains (e.g., Criminal Law and Civil Law), evaluate robustness under adversarial or ambiguous queries, and explore stronger evidence-aware confidence estimators (e.g., entailment-based critics) to further reduce hallucination while minimizing additional latency.

6. Conclusion

This paper presented a self-reflection mechanism for enhancing legal question answering systems. The hybrid scoring formula combining semantic similarity (α=0.4) with sigmoid-normalized confidence (β=0.6) enables effective quality assessment. With quality threshold θ=0.75 and maximum n=3 iterations, the system achieves significant improvements: accuracy increases from 81.5% to 86.7% (+5.2 percentage points) while hallucination rate decreases from 18.7% to 9.3% (-9.4 percentage points).

Comprehensive ablation studies confirm that both scoring components contribute to these improvements, with the hybrid approach outperforming single-metric methods by 5.5-6.9 percentage points. The average 1.8 iterations per query demonstrates that the mechanism efficiently focuses refinement on difficult queries.

Our findings suggest that self-reflection provides an effective approach for improving the reliability of LLM-based legal question answering systems without requiring model retraining. Future work will explore extensions to additional Vietnamese legal domains and integration with multi-turn dialogue systems for handling ambiguous queries.

7. Declarations

7.1. Author Contributions

Thi Vuong Pham: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, Writing – Review & Editing; **Nguyet Minh Phan:** Software, Formal analysis, Validation, Writing – Original Draft; **Bao Quynh Cao:** Investigation, Data Curation, Software; **Cong Phuc Truong:** Software, Validation, Visualization; **Thanh Duy Nguyen:** Investigation, Data Curation; **Minh Vy Tien**: Formal analysis, Visualization; **Ngoc Son Nguyen:** Resources, Investigation.

7.2. Institutional Review Board Statement

Not applicable. This study does not involve human subjects or personal data.

7.3. Informed Consent Statement

Not applicable.

7.4. Data Availability Statement

The datasets generated and analyzed during the current study are not publicly available. Upon reasonable request, we may share a limited set of non-sensitive examples and aggregated statistics that do not include the evaluation set or derived datasets.

The Vietnamese Labor Law documents used in this study are publicly accessible legal documents issued by the National Assembly of Vietnam.

7.5. Acknowledgment

Not applicable.

7.6. Conflicts of Interest

The authors declare no conflicts of interest.

8. References

[1] W. X. Zhao et al., "A Survey of Large Language Models," *arXiv preprint arXiv:2303.18223*, 2023. https://doi.org/10.48550/arXiv.2303.18223.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-T. Yih, T. Rocktäschel, S. Riedel, D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Proceedings of the 34th International Conference on Neural Information Processing System*, pp. 9459–9474, 2020. https://dl.acm.org/doi/abs/10.5555/3495724.3496517.

[3] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," *arXiv preprint* arXiv:2312.10997, 2023. https://doi.org/10.48550/arXiv.2312.10997.

[4] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, W.-T Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781, 2020. https://aclanthology.org/2020.emnlp-main.550.pdf.

[5] C. Yang, R. Xu, L. Luo, and S. Pan, "Knowledge Graph and Large Language Model Co-learning via Structure-Oriented Retrieval-Augmented Generation," *IEEE Data Engineering Bulletin*, vol. 47, no. 1, pp. 9–46, 2024. https://par.nsf.gov/servlets/purl/10590165.

[6] D. V. Dang et al., "Knowledge Graph-Based Legal Query System with LLM and Retrieval Augmented Generation," in *Recent Challenges in Intelligent Information and Database Systems (ACIIDS), Communications in Computer and Information Science*, vol. 2493, pp. 161–172, 2025. https://doi.org/10.1007/978-981-96-5881-7_13.

[7] Z. Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023. https://doi.org/10.1145/3571730.

[8] L. Huang et al., "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *arXiv preprint arXiv:2311.05232*, 2023. https://doi.org/10.48550/arXiv.2311.05232.

[9] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On Faithfulness and Factuality in Abstractive Summarization," in Proc. *of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020. https://doi.org/10.18653/v1/2020.acl-main.173.

[10] N. Shinn, F. Cassano, A. Gopinath, and S. Yao, "Reflexion: Language Agents with Verbal Reinforcement Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 8634–8652, 2023. https://doi.org/10.48550/arXiv.2303.11366.

[11] A. Madaan et al., "Self-Refine: Iterative Refinement with Self-Feedback," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 46512–46594, 2023. https://doi.org/10.48550/arXiv.2303.17651.

[12] Ministry of Justice of Vietnam, "Circular No. 25/2011/TT-BTP on the Format of and Techniques for Presenting Legal Documents of the Government, the Prime Minister, Ministers and Heads of Ministerial-Level Agencies and Joint Legal Documents," Hanoi, Vietnam, 2011. Available: https://thuvienphapluat.vn/van-ban/EN/Linh-vuc-khac/Circular-No-25-2011-TT-BTP-on-the-format-of-and-techniques-for-presenting-legal/136968/tieng-anh.aspx.

[13] National Assembly of Vietnam, "Labor Code No. 45/2019/QH14," Hanoi, Vietnam, 2019. https://english.luatvietnam.vn/labor-code-no-45-2019-qh14-dated-november-20-2019-of-the-national-assembly-179015-doc1.html.

[14] Y. Meng, J. Huang, Y. Zhang, and J. Han, "Generating Training Data with Language Models: Towards Zero-Shot Language Understanding," *arXiv preprint arXiv:2202.04538*, 2022. https://doi.org/10.48550/arXiv.2202.04538.

[15] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009. https://dl.acm.org/doi/abs/10.1561/1500000019.

[16] J. J. Jiang and D. W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," in *Proc. 10th Research on Computational Linguistics International Conference (ROCLING X)*, pp. 19–33, 1997. https://arxiv.org/abs/cmp-lg/9709008.

[17] J. Metcalfe and A. P. Shimamura, *Metacognition: Knowing about Knowing*. Cambridge, MA, USA: MIT Press, 1994. https://doi.org/10.7551/mitpress/4561.001.0001.

[18] X. Wang et al., "Self-Consistency Improves Chain-of-Thought Reasoning in Language Models," in *arXiv preprint arXiv:2203.11171*, 2023. https://doi.org/10.48550/arXiv.2203.11171.

[19] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *arXiv preprint arXiv:2201.11903* , 2022. https://doi.org/10.48550/arXiv.2201.11903.

[20] S. Kadavath et al., "Language Models (Mostly) Know What They Know," *arXiv preprint arXiv:2207.05221*, 2022. https://doi.org/10.48550/arXiv.2207.05221.

[21] S. Chen et al., "FELM: Benchmarking Factuality Evaluation of Large Language Models," *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, vol. 36, 2023. https://doi.org/10.48550/arXiv.2310.00741.

[22] Y. Du et al., "Improving Factuality and Reasoning in Language Models through Multiagent Debate," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, PMLR, vol. 235, pp. 11733–11763, 2024. https://proceedings.mlr.press/v235/du24e.html.

[23] L. Pan et al., "Fact-Checking Complex Claims with Program-Guided Reasoning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6981–7004, 2023. https://doi.org/10.18653/v1/2023.acl-long.386.

[24] J. Huang and K. C. Chang, "Towards Reasoning in Large Language Models: A Survey," *Findings of the Association for Computational Linguistics (ACL)*, pp. 1049–1065, 2023. https://doi.org/10.18653/v1/2023.findings-acl.67.

[25] H. Q. Ngo, H. D. Nguyen, and N.-A. Le-Khac, "Ontology Knowledge Map Approach Towards Building Linked Data for Vietnamese Legal Applications," *Vietnam Journal of Computer Science*, vol. 11, no. 2, pp. 323–342, 2024. https://doi.org/10.1142/S2196888824500015.

[26] V. T. Pham, H. D. Nguyen, T. Le, B. Nguyen, and H. Q. Ngo, "Ontology-Based Solution for Building an Intelligent Searching System on Traffic Law Documents," *in Proc. International Conference on Agents and Artificial Intelligence (ICAART)*, vol. 1, pp. 217–224, 2023. https://doi.org/10.5220/0011635500003393.

[27] H. Nguyen et al., "Intelligent Search System for Resume and Labor Law Using Knowledge Graphs," *PeerJ Computer Science*, vol. 10, e1786, 2024. https://doi.org/10.7717/peerj-cs.1786.

[28] D. V. Dang, V. T. Pham, T. Cao, N. Do, H. Q. Ngo, and H. D. Nguyen, "A Practical Approach to Leverage Knowledge Graphs for Legal Query," in *Intelligent Systems and Data Science*, vol. 1949, Springer, pp. 271–284, 2024. https://doi.org/10.1007/978-981-99-7649-2_21.

[29] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained Language Models for Vietnamese," in *Findings of the Association for Computational Linguistics*: EMNLP, pp. 1037–1042, 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.92.

[30] T. Vu et al., "VnCoreNLP: A Vietnamese Natural Language Processing Toolkit," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*: Demonstrations, pp. 56–60, 2018. https://doi.org/10.18653/v1/N18-5012.

[31] N. Minh et al., "ViHealthBERT: Pre-trained Language Models for Vietnamese in Health Text Mining," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 328–337, 2022. https://aclanthology.org/2022.lrec-1.35/.

[32] A. Conneau et al., "Unsupervised Cross-Lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020. https://doi.org/10.48550/arXiv.1911.02116.

[33] I. Chalkidis et al., "FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing," in *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. https://doi.org/10.18653/v1/2022.acl-long.301.

[34] Y. Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," *arXiv preprint arXiv:2309.01219*, 2023. https://doi.org/10.48550/arXiv.2309.01219.

[35] P. Manakul, A. Liusie, and M. J. F. Gales, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. https://doi.org/10.18653/v1/2023.emnlp-main.557.

[36] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han, "Towards a Unified Multi-Dimensional Evaluator for Text Generation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2023–2038, 2022. https://doi.org/10.18653/v1/2022.emnlp-main.131.

[37] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. https://doi.org/10.18653/v1/D19-1410.

[38] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013. https://doi.org/10.48550/arXiv.1301.3781.

[39] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. https://books.google.com/books/about/Deep_Learning.html?id=Np9SDQAAQBAJ.

[40] H. Liu et al., "Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 19587–19599,2022. https://doi.org/10.48550/arXiv.2205.05638.

[41] C. Gao et al., "StrategyLLM: Large Language Models as Strategy Generators, Executors, Optimizers, and Evaluators for Problem Solving," *arXiv preprint arXiv:2311.08803*, 2023. https://doi.org/10.48550/arXiv.2311.08803.

[42] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008. Available: https://books.google.com/books/about/Introduction_to_Information_Retrieval.html?id=t1PoSh4uwVcC.

[43] S. E. Robertson, "The Probability Ranking Principle in IR," *Journal of Documentation*, vol. 33, no. 4, pp. 294–304, 1977. https://doi.org/10.1108/eb026647.