A Benchmark for Structured Multihop Reasoning in Cross-Institution University Admission Advisory Question Answering

Long S. T. Nguyen[1,2], Tin T. Ngo[1], Dung N. H. Le[1], Quynh T. N. Vo[1], Dung H. Nguyen[1], Khang N. Le[1], and Tho T. Quan[1,2(✉)]

[1] URA Research Group, Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam
qttho@hcmut.edu.vn

[2] Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam

Abstract. University admission advisory is a high-stakes decision-making process in which applicants must interpret and reconcile heterogeneous institutional regulations, eligibility criteria, tuition policies, scholarships, deadlines, and procedures across multiple universities. While recent advances in Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) enable document-grounded question answering, existing QA benchmarks are largely built on homogeneous corpora and seldom capture the cross-institution and regulation-driven multihop reasoning required in real admission scenarios, especially for Vietnamese. We present VIXSTORY (Vietnamese Cross-Institution University Admission Advisory QA), a benchmark designed to evaluate hierarchical and graph-structured reasoning under both intra-institution and cross-institution settings. VIXSTORY is constructed via a controlled semi-automatic pipeline that combines document segmentation, embedding-based topic clustering within institutions, cross-institution topic alignment, and LLM-assisted question generation with strict post-validation for faithfulness. We further provide a comprehensive evaluation of representative retrieval and reasoning paradigms, from lexical/dense baselines to iterative multihop and graph-based RAG methods. Results show substantial gaps between flat retrieval and structured approaches, with graph-based methods achieving markedly stronger performance and robustness as hop depth increases, highlighting the importance of explicit structure for realistic regulation-driven advisory systems. For reproducibility and future research, VIXSTORY is publicly released at https://huggingface.co/datasets/ura-hcmut/VixSTORY.

Keywords: Vietnamese question answering · university admission advisory · regulation-centric QA · multihop reasoning

1 Introduction

University admission advisory is a complex, high-stakes decision-making process that requires prospective students to reason over heterogeneous institutional reg-

C. Long et al. (Eds.): PAKDD 2026, LNAI 16603, pp. 235–247, 2027.
https://doi.org/10.1007/978-981-92-2014-4_19

ulations, eligibility criteria, tuition policies, scholarships, deadlines, and procedural requirements across universities [14,16]. Unlike factual information lookup, advisory queries demand the reconciliation of regulatory constraints to support comparison and informed decisions. In practice, applicants must integrate evidence from multiple documents within and across institutions [4,10]. In Vietnam, this challenge is amplified by the absence of standardized policy representations, as regulations vary substantially in terminology, structure, and style across institutions [1,13].

Recent advances in *Large Language Models* (LLMs) and *Retrieval-Augmented Generation* (RAG) enable document-grounded question answering [2,11]. However, most *Question Answering* (QA) benchmarks assume a single corpus with relatively homogeneous evidence. Although some datasets support multihop reasoning within one institution, they do not capture the additional complexity of *cross-institution advisory reasoning* [4,13,15]. In cross-institution settings, systems must align related regulations, resolve subtle policy differences, and integrate heterogeneous evidence. These challenges expose a deeper modeling limitation: flat retrieval pipelines lack explicit mechanisms to represent structured relationships among regulations spanning multiple institutions.

To address this gap, we introduce ***Vi****etnamese* ***Cross****-In****S****titution Universi****T****y Admissi****O****n Adviso****RY*** *QA* (VIXSTORY), a benchmark for regulation-driven university admission advisory. VIXSTORY supports both intra-institution and cross-institution multihop reasoning, where answering a query requires synthesizing evidence from multiple documents and reconciling institutional constraints. From an autonomous intelligence perspective, this task demands structured internal representations of a complex regulatory environment to support multi-step reasoning and decision-making.

A central objective of VIXSTORY is to evaluate *structured multihop reasoning* as a form of hierarchical representation. Admission regulations naturally form layered abstractions, from clause-level rules to institution-level policies. Effective advisory systems must therefore reason across multiple levels while modeling cross-document and cross-institution relations. VIXSTORY enables systematic comparison between flat RAG pipelines and structured approaches, including hierarchical and graph-based reasoning. The dataset is constructed through a controlled semi-automatic process with strict post-validation to ensure faithful grounding, genuine multihop integration, and diverse advisory intents.

Using VIXSTORY, we benchmark representative retrieval and reasoning paradigms, from lexical and dense baselines to advanced structured RAG variants [5–7,11,20]. Results reveal substantial performance gaps between flat and structured approaches, particularly in cross-institution scenarios, underscoring the necessity of explicit structure and multihop planning in regulation-driven advisory tasks.

Our contributions are summarized as follows.

– We formalize university admission advisory question answering as a regulation-centric structured multihop reasoning task encompassing both intra-institution and cross-institution scenarios.

- We introduce VIXSTORY, a benchmark dataset designed to evaluate hierarchical representations and structured reasoning in realistic Vietnamese university admission advisory settings.
- We provide a systematic empirical evaluation of existing RAG-based methods, offering diagnostic insights for the development of future autonomous advisory systems.

2 Related Work

2.1 Vietnamese Regulation-Centric and Advisory QA Datasets

Research on Vietnamese question answering over institutional regulations remains relatively limited, with most existing datasets concentrated in the legal and educational domains. Prior work has introduced regulation-centric QA resources for Vietnamese statute laws [3], educational management policies [13], bidding regulations [8], university training rules [4], and labor law documents [17]. While these datasets provide important foundations for regulation-oriented QA, they are typically constructed within a single institution or document collection and predominantly emphasize factoid retrieval or shallow comprehension. UniQA [15] takes a step toward advisory question answering for university admissions by highlighting regulation-driven reasoning in educational decision support. However, it remains restricted to a single institution and does not require systems to reconcile heterogeneous admission policies across universities. As a result, existing Vietnamese regulation-centric and advisory QA datasets do not support cross-document and cross-institution multihop reasoning, leaving an open gap in evaluating advisory systems under realistic admission scenarios where institutionally independent evidence sources must be jointly considered. VIXSTORY is designed to fill this gap.

2.2 Benchmarks for Hierarchical and Multihop Reasoning

Multihop question answering benchmarks have been widely studied as a means to evaluate systems that must aggregate evidence across multiple documents rather than rely on single-passage lookup. Seminal datasets such as HotpotQA [21], along with later compositional benchmarks including MuSiQue [19] and 2WikiMultiHopQA [9], require multi-step inference over disjoint passages and have driven advances in iterative retrieval and reasoning. For Vietnamese, ViMQA [10] extends this paradigm to a low-resource setting using Wikipedia-based multihop questions. Despite their impact, these benchmarks are predominantly constructed over homogeneous corpora and focus on general-purpose knowledge. They do not explicitly model structured regulatory hierarchies, institution-specific constraints, or cross-institution policy variability, and therefore do not evaluate a system's ability to perform hierarchical aggregation and reconcile heterogeneous regulations across institutionally independent documents. This limitation motivates benchmarks such as VIXSTORY, which are explicitly designed to evaluate structured multihop reasoning under regulation-driven, cross-institution advisory settings.

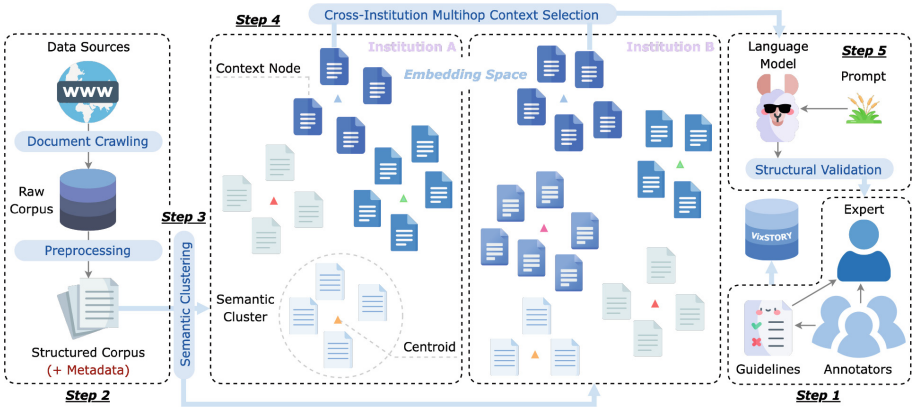

Fig. 1. VIXSTORY dataset construction pipeline for cross-institution multihop university admission advisory QA.

3 VIXSTORY Dataset

3.1 Dataset Construction

We construct VIXSTORY via a controlled semi-automatic pipeline that combines expert guidance, semantic structuring, cross-institution context selection, and language-model-based generation to support structured multihop advisory reasoning for Vietnamese university admissions, as illustrated in Fig. 1.

Step 1: Guidelines and Expert Supervision. We begin by defining expert-curated *guidelines* that formalize valid university admission advisory behavior, including admissible question types, evidence grounding requirements, and multihop reasoning constraints. A small group of *annotators*, together with a domain *expert*, refines these guidelines to ensure regulatory plausibility and decision relevance. Formally, these guidelines define a constraint set $\mathcal{G}$ over generated instances, covering structural validity, evidence completeness, and cross-institution consistency.

Step 2: Corpus Construction from Data Sources. Admission-related regulations are collected from official university websites via automated *document crawling*, forming a *raw corpus*. Through *preprocessing*, documents are normalized, de-noised, and organized into a *structured corpus* with associated *metadata*. Each resulting context is treated as an atomic regulatory unit represented as a tuple $\langle c_i, s_i \rangle$, where c_i denotes the textual content and s_i is a stable institution identifier. This yields a corpus $\mathcal{C} = \{(c_i, s_i)\}$ that explicitly preserves institutional provenance for downstream reasoning.

Step 3: Semantic Clustering within Institutions. Each context is embedded into a shared *embedding space* as a *context node* with vector representation $\mathbf{e}_i \in \mathbb{R}^d$, which is ℓ_2-normalized to $\tilde{\mathbf{e}}_i = \mathbf{e}_i / \|\mathbf{e}_i\|_2$. To capture topical structure while respecting institutional boundaries, we perform *semantic clustering*

Table 1. Statistics of the VIXSTORY test split by hop level, type, and scope. *NoS* denotes the number of samples. Lengths are reported as *min/mean/max*.

Category	NoS	Vocab	Question length	Answer length	Context length
By number of hops					
1	222	2542	54/148.9/277	86/269.9/798	482/1396.8/3923
2	521	4702	92/149.7/335	134/367.4/985	1186/2689.2/9447
3	492	6481	91/183.5/370	96/427.8/1019	1036/5186.3/27272
4	227	4525	97/206.6/528	190/444.0/963	1987/4898.3/12525
5	238	4781	96/234.5/528	183/456.8/772	1703/5928.4/10609
By question type					
Comparative	921	7894	91/150.6/509	136/430.1/1019	1036/4279.3/27272
Procedural	307	3761	54/202.9/476	105/372.9/617	636/4166.7/12525
Conditional	250	3502	90/246.2/528	86/336.6/643	704/3473.2/7232
Verification	47	2480	97/154.4/274	142/427.1/768	2203/4212.1/7705
Factual	167	3113	60/198.0/382	96/320.0/588	482/2788.3/11156
Bridge	8	1138	94/147.4/194	228/380.4/480	2997/4162.0/6830
By reasoning scope					
Cross-institution	976	8219	91/150.8/509	136/429.5/1019	1036/4275.1/27272
Within-institution	724	4053	54/216.7/528	86/348.2/643	482/3609.3/12525
All	1700	8484	54/178.8/528	86/394.9/1019	482/3991.6/27272

independently for each institution s. For each institution-specific cluster $C_k^{(s)}$, we compute a centroid $\mathbf{c}_k^{(s)} = \frac{1}{|C_k^{(s)}|} \sum_{\tilde{\mathbf{e}}_i \in C_k^{(s)}} \tilde{\mathbf{e}}_i$ and rank member contexts by Euclidean distance $\|\tilde{\mathbf{e}}_i - \mathbf{c}_k^{(s)}\|_2$. The nearest contexts are retained as representatives, yielding institution-scoped *semantic clusters* that approximate latent admission topics.

Step 4: Cross-Institution Multihop Context Selection. To support comparative and decision-oriented reasoning, we align semantically related clusters across universities. Clusters $C_k^{(s)}$ and $C_{k'}^{(s')}$ with $s \neq s'$ are linked if $\cos(\mathbf{c}_k^{(s)}, \mathbf{c}_{k'}^{(s')}) \geq \tau$, inducing a lightweight cross-institution semantic structure. Given a hop level h, we sample multihop context tuples $\mathcal{T} = \{c_1, \ldots, c_h\}$ either within a single institution or across multiple institutions, subject to overlap and coherence constraints. In particular, tuples are required to satisfy the average semantic relatedness condition $\frac{2}{h(h-1)} \sum_{i<j} \cos(\tilde{\mathbf{e}}_i, \tilde{\mathbf{e}}_j) \geq \tau_{\text{avg}}$, ensuring that each hop contributes meaningfully to a coherent multistep reasoning chain.

Step 5: Language Model Generation and Validation. Each sampled tuple $\mathcal{T}$ is provided to a *language model* through a structured *prompt* to generate advisory question–answer pairs. The model is required to integrate evidence from all h selected contexts, and cross-institution instances must explicitly ref-

Table 2. A representative cross-institution multihop QA instance from VIXSTORY.

Field	Content
Query	Đại học Kinh tế TP. Hồ Chí Minh và Trường Đại học Bách khoa có những tiêu chí học tập nào để xét cấp học bổng cho sinh viên? **English:** What academic criteria do the University of Economics Ho Chi Minh City and Ho Chi Minh City University of Technology use to award scholarships to students?
Answer	Đại học Kinh tế TP. Hồ Chí Minh xét học bổng khuyến khích học tập cho tân sinh viên dựa trên điểm xét tuyển đầu vào (không tính điểm ưu tiên), và đối với sinh viên đang học thì căn cứ vào kết quả học tập và rèn luyện. Trong khi đó, Trường Đại học Bách khoa yêu cầu sinh viên có học lực từ 7.00 trở lên, đồng thời có thể xét thêm các tiêu chí như năng lực tiếng Anh hoặc thành tích nghiên cứu khoa học. **English:** UEH awards scholarships based on entrance examination scores for freshmen and academic and conduct performance for enrolled students, whereas HCMUT requires a minimum GPA of 7.00 and may consider additional criteria such as English proficiency or research achievements.
Number of hops	2
Reasoning scope	cross-institution
Topic	comparative
School pair	University of Economics Ho Chi Minh City (UEH), Vietnam ↔ Ho Chi Minh City University of Technology (HCMUT), Vietnam
Evidence	`context_1`: Quy định về tiêu chí xét học bổng khuyến khích học tập của Đại học Kinh tế TP. Hồ Chí Minh dựa trên điểm xét tuyển và kết quả học tập, rèn luyện. `context_2`: Quy định của Trường Đại học Bách khoa về điều kiện học lực tối thiểu (GPA $\geqslant$ 7.00) và các tiêu chí bổ sung để xét học bổng. **English:** The evidence consists of scholarship regulations from UEH and HCMUT specifying different academic criteria.
Reasoning	Câu trả lời được hình thành bằng cách so sánh trực tiếp các tiêu chí xét học bổng được quy định độc lập bởi hai trường đại học. `context_1` cung cấp tiêu chí của UEH, trong khi `context_2` mô tả điều kiện của HCMUT. Việc kết hợp cả hai nguồn là cần thiết để trả lời đầy đủ câu hỏi mang tính so sánh liên trường. **English:** The answer compares independently defined scholarship criteria from UEH and HCMUT, requiring evidence from both institutions to support a cross-institution advisory decision.

erence multiple institutions. To mitigate hallucination, the model is instructed to produce an intermediate reasoning trace, enforcing explicit evidence integration across all selected contexts. Generated outputs undergo *structural validation* against the constraint set $\mathcal{G}$, verifying multihop completeness, evidence coverage, and logical consistency, with expert review applied to ambiguous cases. All validated instances are aggregated to form the final VIXSTORY benchmark dataset.

3.2 Dataset Statistics

Table 1 summarizes the statistics of the VIXSTORY test split, drawn from a full dataset of over 3,000 instances, across hop levels, question types, and reasoning scopes. The test split covers five hop levels, with most samples concentrated in 2-hop and 3-hop settings, reflecting the prevalence of moderate multistep advisory reasoning. Across question types, *comparative* questions dominate and involve the largest vocabularies and context lengths, while *verification* and *procedural* questions also exhibit high mean context lengths due to the need for evidence-intensive rule checking and multi-step admission workflows. *Cross-institution* samples account for a substantial portion of the split and consistently exhibit larger vocabularies and longer contexts than *within-institution* ones, indicating the added difficulty of aligning institutionally independent regulations. Overall, these statistics confirm that VIXSTORY is a challenging benchmark for structured multihop reasoning beyond flat retrieval-based approaches.

3.3 Example Instances

Table 2 presents a representative example from VIXSTORY, including the question, answer, hop level, reasoning scope, and supporting evidence. English paraphrases are provided only for readability.

4 Experiments

We assess system performance along multiple complementary dimensions, including answer correctness, multihop reasoning quality, retrieval effectiveness, inference latency, and the ability to integrate evidence across heterogeneous institutions. All methods are compared against representative baseline approaches under a unified evaluation protocol.

4.1 Dataset

All experiments are conducted on the test split of VIXSTORY, which is specifically designed to assess multihop advisory reasoning over Vietnamese university admission regulations. The test set spans diverse hop levels and advisory reasoning patterns, including cross-document evidence integration, policy comparison across institutions, and decision-oriented synthesis under heterogeneous admission rules.

4.2 Baselines

We compare representative baselines on the VIXSTORY benchmark, drawn from three major families of retrieval-augmented and multihop question answering approaches.

Naive RAG. We consider retrieval-augmented generation pipelines based on flat passage retrieval, including: (1) lexical retrieval using BM25, (2) dense retrieval via embedding similarity, and (3) hybrid retrieval combining lexical and dense scores through weighted interpolation. These baselines represent RAG settings without explicit multihop reasoning or structural awareness.

Reasoning-Guided Multihop QA. IRCoT [20] is a multistep QA approach that interleaves retrieval with *Chain-of-Thought* reasoning to iteratively guide evidence selection using a strong language model, without explicitly modeling institutional structure or policy-level dependencies.

Graph-Based RAG Methods. We include several representative graph-aware retrieval frameworks that incorporate graph structures into indexing and retrieval to support multihop reasoning:

- **RAPTOR** [18] constructs a hierarchical tree of textual units via recursive summarization, enabling retrieval across multiple levels of abstraction.
- **MiniRAG** [5] is a lightweight graph-based system that combines a small language model with a heterogeneous graph index to achieve efficient structured retrieval under limited computational budgets.
- **LightRAG** [6] emphasizes efficient graph construction and traversal strategies to balance retrieval quality and inference latency.
- **HippoRAG 2** [7] draws inspiration from hippocampal memory mechanisms, modeling long-term knowledge as a graph to improve recall and evidence integration across distant reasoning hops.

Table 3. Overall performance on the VIXSTORY test set. **Bold** and underline indicate the best and second-best results, respectively.

	Quality				*Efficiency*		*Graph Structure*		*Cost*
Method	F1	Recall@5	Recall@10	Judge	Avg. Time	Avg. Tokens	Nodes	Edges	Graph Tokens
BM25	0.268	0.204	0.248	0.213	5.96	1472.9	–	–	–
Dense	0.318	0.240	0.294	0.235	8.44	1654.0	–	–	–
Hybrid	0.360	0.298	0.315	0.296	11.90	1873.6	–	–	–
IRCoT	0.366	0.355	0.441	0.342	16.02	2102.6	–	–	–
MiniRAG	0.476	0.419	0.487	0.400	17.46	2428.9	4987	1489	4,129,189
RAPTOR	0.509	0.479	0.512	0.444	19.05	2639.1	5161	2177	4,486,537
LightRAG	0.639	0.537	0.564	0.603	22.80	2805.0	6873	7231	4,768,451
HippoRAG	**0.656**	**0.614**	**0.697**	**0.644**	24.56	2959.4	7429	9639	5,030,919

4.3 Evaluation Metrics

We evaluate system performance at both the answer and retrieval levels. For answer evaluation, we report token-level $F1$ to measure surface-level overlap between predicted and reference answers. Because $F1$ alone is insufficient to capture advisory correctness and policy consistency in multihop admission QA, we additionally adopt an *LLM-as-a-Judge* protocol [12], in which a strong external language model assesses correctness, completeness, and consistency with supporting evidence. For retrieval, we report Recall@k, which measures the proportion of gold supporting contexts recovered within the top-k retrieved results.

4.4 Experimental Setup

To ensure fair comparison and reproducibility, all systems share the same language model backbone and evaluation protocol. We adopt `GPT-4o-mini`[1] as the unified language model for answer generation, chosen for its balance between reasoning capability and computational efficiency. Text embeddings are generated using `text-embedding-3-small`[2], a multilingual embedding model supporting Vietnamese. All baselines are evaluated using default hyperparameters to reflect realistic out-of-the-box performance. LLM-as-a-Judge evaluations are conducted using `GPT-4o`[3] under a standardized rubric to ensure consistent and comparable scoring across systems.

4.5 Results and Analysis

Overall Results. Table 3 shows that flat RAG methods perform poorly ($F1 \leqslant 0.360$), and higher recall alone is insufficient: IRCoT achieves the best non-graph Recall@10 (0.441) yet only Judge = 0.342, indicating that iterative reasoning without structural awareness cannot translate retrieved evidence into correct answers. Graph-based methods dominate: HippoRAG leads with $F1 = 0.656$, Recall@10 = 0.697, and Judge = 0.644, followed closely by LightRAG (0.639, 0.564, 0.603), confirming that explicit structure is critical for both retrieval and reasoning in cross-institution advisory QA.

Hop-Level Robustness. Table 4 reveals a steep degradation with hop depth: the aggregate failure rate (averaged across all M=8 methods per sample) escalates from 15.3% (hop 1) → 38.9% → 63.9% → 86.3% → 94.1% (hop 5), reflecting compounding decomposition and integration errors rather than isolated retrieval misses. BM25 fails on 27.5% of hop-1 and 100% of hop-5 queries; HippoRAG fails on only 6.3% and 92.0% respectively, retaining 69% of its hop-1 Judge (0.779 → 0.539). LightRAG behaves similarly (0.663 → 0.484, 73% retention) and even scores higher at hop 2 (0.670) than hop 1 (0.663), suggesting that its graph traversal is particularly effective for moderate-depth queries where cross-document linking provides the greatest advantage.

[1] https://developers.openai.com/api/docs/models/gpt-4o-mini.
[2] https://developers.openai.com/api/docs/models/text-embedding-3-small.
[3] https://developers.openai.com/api/docs/models/gpt-4o.

Table 4. Hop-wise Judge scores on VIXSTORY. **Bold** and underline indicate the best and second-best results, respectively.

Method	Hop 1	Hop 2	Hop 3	Hop 4	Hop 5	All
BM25	0.394	0.238	0.203	0.138	0.078	0.213
Dense	0.420	0.257	0.234	0.147	0.096	0.235
Hybrid	0.452	0.345	0.283	0.212	0.149	0.296
IRCoT	0.508	0.379	0.327	0.263	0.216	0.342
MiniRAG	0.575	0.432	0.381	0.325	0.275	0.400
RAPTOR	0.644	0.475	0.421	0.345	0.327	0.444
LightRAG	0.663	0.670	0.591	0.537	0.484	0.603
HippoRAG	**0.779**	**0.686**	**0.632**	**0.551**	**0.539**	**0.644**

Error Analysis By Question Type and Reasoning Scope. Failure rates differ sharply by question type: *comparative* 69.0%, *bridge* 64.1%, *verification* 60.1%, *procedural* 47.0%, *conditional* 41.0%, *factual* 33.1%. The comparative–factual gap is highly significant (Welch's t-test, $p = 5.43\times10^{-146}$), confirming that multi-entity policy alignment, not fact lookup, is the dominant failure mode. Cross-institution samples exhibit substantially higher failure rates than within-institution ones (68.5% vs. 41.7%, $\Delta = 26.8$ pp; Welch's $t = 5.92$, $p = 5.98\times10^{-9}$), underscoring the difficulty of reconciling heterogeneous regulations across universities.

Length Bias and Context Complexity. We define per-sample difficulty as $d_i = \frac{1}{M}\sum_{m=1}^{M} \mathbf{1}[\text{method } m \text{ fails on sample } i]$, i.e., the fraction of methods that answer sample i incorrectly. Difficulty correlates strongly with context character length (Pearson's $r = 0.749$, $p = 2.52\times10^{-217}$), context token count ($r = 0.750$, $p = 4.23\times10^{-181}$), and reference answer length ($r = 0.632$, $p = 4.64\times10^{-100}$). The longest-context quartile fails at 70.3% versus 55.3% for the shortest (Welch's t-test, $p = 1.95\times10^{-95}$). Moreover, failed outputs have a lower generated-to-reference length ratio than passed ones (0.553 vs. 0.703), indicating that heavier reasoning load leads to truncated, incomplete synthesis.

Efficiency Trade-offs. Graph methods incur higher costs—HippoRAG averages 24.56s and 2,959.4 tokens per query versus 5.96s and 1,472.9 tokens for BM25—yet deliver substantially stronger quality: HippoRAG's Judge (0.644) is three times that of BM25 (0.213), a favorable trade-off for high-stakes advisory applications where correctness outweighs latency.

Graph Structure and Indexing Cost. Performance scales with relational density rather than graph size alone: the edge-to-node ratio rises from 0.30 (MiniRAG) to 1.30 (HippoRAG), closely tracking Judge gains (0.400 → 0.644). The graph token cost increases by only 22% (∼4.1M → ∼5.0M) while Judge improves by 61%, confirming favorable cost–quality scaling for graph-based indexing.

Summary. VIXSTORY effectively discriminates flat, reasoning-guided, and graph-based approaches. Graph methods excel especially at higher hops and

cross-institution settings, while systematic gaps across question types, scopes, and context lengths confirm the necessity of explicit structural reasoning for regulation-driven advisory QA.

5 Conclusion

We introduced VixSTORY, a Vietnamese benchmark for regulation-driven university admission advisory QA that targets both intra-institution and cross-institution multihop reasoning. Constructed through a controlled semi-automatic pipeline with semantic structuring and strict post-validation, VixSTORY provides a realistic and challenging setting for evaluating structured reasoning over heterogeneous institutional regulations. Our experiments show that graph-based methods consistently outperform flat and reasoning-guided baselines, with the advantage becoming more evident as reasoning complexity increases. These findings highlight the importance of explicit structural modeling for reliable advisory QA and position VixSTORY as a useful benchmark for future research on robust, regulation-grounded question answering systems.

References

1. Bui, T., et al.: Cross-data knowledge graph construction for LLM-enabled educational question-answering system: a case study at HCMUT. In: Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia, AIQAM '24, pp. 36–43. Association for Computing Machinery, New York (2024). https://doi.org/10.1145/3643479.3662055
2. Chang, Y., et al.: A Survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol. **15**(3) (2024). https://doi.org/10.1145/3641289
3. Do, D.T., et al.: A summary of the ALQAC 2024 competition. In: 2024 16th International Conference on Knowledge and System Engineering (KSE), pp. 422–427 (2024). https://doi.org/10.1109/KSE63888.2024.11063484
4. Do, T.P.P., Cao, N.D.D., Tran, K.Q., Van Nguyen, K.: R2GQA: retriever-reader-generator question answering system to support students understanding legal regulations in higher education. Artif. Intell. Law (2025)
5. Fan, T., Wang, J., Ren, X., Huang, C.: MiniRAG: towards extremely simple retrieval-augmented generation (2025)
6. Guo, Z., Xia, L., Yu, Y., Ao, T., Huang, C.: LightRAG: simple and fast retrieval-augmented generation. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 10746–10761. Association for Computational Linguistics, Suzhou (2025).https://doi.org/10.18653/v1/2025.findings-emnlp.568
7. Gutiérrez, B.J., Shu, Y., Qi, W., Zhou, S., Su, Y.: From RAG to memory: non-parametric continual learning for large language models. In: Forty-second International Conference on Machine Learning (2025). https://openreview.net/forum?id=LWH8yn4HS2
8. Ha, N.T., et al.: Vietnamese legal question answering: an experimental study. In: 2024 16th International Conference on Knowledge and System Engineering (KSE), pp. 440–446 (2024). https://doi.org/10.1109/KSE63888.2024.11063637

9. Ho, X., Duong Nguyen, A.K., Sugawara, S., Aizawa, A.: Constructing a Multi-hop QA dataset for comprehensive evaluation of reasoning steps. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics, pp. 6609–6625. International Committee on Computational Linguistics, Barcelona (Online) (2020).https://doi.org/10.18653/v1/2020.coling-main.580
10. Le, K., Nguyen, H., Le Thanh, T., Nguyen, M.: VIMQA: a Vietnamese dataset for advanced reasoning and explainable multi-hop question answering. In: Calzolari, N., Béchet, F., et al. (eds.) Proceedings of the Thirteenth Language Resources and Evaluation Conference, pp. 6521–6529. European Language Resources Association, Marseille (2022). https://aclanthology.org/2022.lrec-1.700/
11. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20. Curran Associates Inc., Red Hook (2020)
12. Li, D., et al.: From generation to judgment: opportunities and challenges of LLM-as-a-judge. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 2757–2791. Association for Computational Linguistics, Suzhou (2025). https://doi.org/10.18653/v1/2025.emnlp-main.138
13. Minh, D.D., Van, V.N., Cong, T.D.: Using large language models for education managements in Vietnamese with low resources. In: Oco, N., Dita, S.N., Borlongan, A.M., Kim, J.B. (eds.) Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation, pp. 20–34. Tokyo University of Foreign Studies, Tokyo (2024). https://aclanthology.org/2024.paclic-1.3/
14. Ng, P., Lee, D., Wong, P., Lam, R.: Making a higher education institution choice: differences in the susceptibility to online information on students' advice-seeking behavior. Online Inf. Rev. **44**(4), 847–861 (2020)
15. Nguyen, L., Vo, Q., Luu, H., Quan, T.: When in doubt, ask first: a unified retrieval agent-based system for ambiguous and unanswerable question answering. In: Inui, K., Sakti, S., Wang, H., Wong, D.F., Bhattacharyya, P., Banerjee, B., Ekbal, A., Chakraborty, T., Singh, D.P. (eds.) Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pp. 452–472. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Mumbai (2025). https://aclanthology.org/2025.findings-ijcnlp.27/
16. Nguyen, L.S.T., Quan, T.T.: URAG: implementing a unified hybrid RAG for precise answers in university admission chatbots – a case study at HCMUT. In: Buntine, W., Fjeld, M., Tran, T., Tran, M.T., Huynh Thi Thanh, B., Miyoshi, T. (eds.) Information and Communication Technology, pp. 82–93. Springer Nature Singapore, Singapore (2025)
17. Pham, V., Le, H.H., Ngo, T.P., Nguyen, B., Nguyen, D., Nguyen, H.D.: Enhancing legal research through knowledge-infused information retrieval for Vietnamese labor law. IAES Int. J. Artif. Intell. (IJ-AI) **13**(4), 3962–3973 (2024). https://doi.org/10.11591/ijai.v13.i4.pp3962-3973
18. Sarthi, P., Abdullah, S., Tuli, A., Khanna, S., Goldie, A., Manning, C.D.: RAPTOR: recursive abstractive processing for tree-organized retrieval. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=GN921JHCRw

19. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: MuSiQue: multihop questions via single-hop question composition. Trans. Assoc. Comput. Linguist. **10**, 539–554 (2022)
20. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 10014–10037. Association for Computational Linguistics, Toronto (2023). https://doi.org/10.18653/v1/2023.acl-long.557
21. Yang, Z., et al.: HotpotQA: a dataset for diverse, explainable multi-hop question answering. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2369–2380. Association for Computational Linguistics, Brussels (2018). https://doi.org/10.18653/v1/D18-1259