Which Works Best for Vietnamese? A Practical Study of Information Retrieval Methods across Domains

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved remarkable progress, yet their reliance on parametric knowledge often leads to hallucinations. Retrieval-Augmented Generation (RAG) mitigates this issue by grounding outputs in external documents, where the quality of retrieval is critical. While retrieval methods have been widely benchmarked in English, it remains unclear which approaches are most effective for Vietnamese, a language characterized by informal queries, noisy documents, and limited resources. Prior studies are restricted to clean datasets or narrow domains, leaving fragmented insights. To the best of our knowledge, we present the first comprehensive benchmark of retrieval methods for Vietnamese across multiple real-world domains. We systematically compare lexical, dense, and hybrid methods on datasets spanning education, legal, healthcare, customer support, lifestyle, and Wikipedia, and introduce two new datasets capturing authentic educational counseling and customer service interactions. Beyond reporting benchmark numbers, we distill a set of empirical insights that clarify trade-offs, highlight domain-specific challenges, and provide practical guidance for building robust Vietnamese QA systems. Together, these contributions offer the first large-scale, practice-oriented perspective on Vietnamese retrieval and inform both academic research and real-world deployment in low-resource languages. All datasets and evaluation scripts are available at https://anonymous.4open.science/r/ViRE.

1 Introduction

Large Language Models (LLMs) have rapidly advanced and now serve as the backbone of many applications, ranging from conversational agents to domain-specific *Question Answering* (QA) systems (Kamalloo et al., 2023; Chang et al., 2024). Despite their impressive fluency and generalization ability, LLMs remain fundamentally limited by their reliance on parametric knowledge (Chang et al., 2024). This limitation often manifests in hallucinations, where the model produces plausible but incorrect content. To mitigate this issue, *Retrieval-Augmented Generation* (RAG) has emerged as a practical paradigm and is increasingly recognized as indispensable for real-world QA systems (Lewis et al., 2020). In RAG, LLMs are supplemented with relevant external documents retrieved from a knowledge base, grounding their responses in factual evidence rather than internal memory alone.

At the heart of every RAG system is the retriever (Fan et al., 2024; Arslan et al., 2024). The quality of retrieval directly governs the accuracy and reliability of generated answers. A wide spectrum of retrieval methods exists, ranging from traditional lexical approaches such as TF–IDF and BM25 to modern dense embedding models and hybrid combinations (Fan et al., 2024). These methods have been extensively benchmarked in English and other high-resource languages, yet the situation is markedly different for Vietnamese (Nguyen and Quan, 2025). Queries in realistic usage scenarios often contain informal expressions, slang, abbreviations, or misspellings. Documents, particularly those from customer support or online forums, are noisy and heterogeneous. Unlike curated corpora such as Wikipedia, which dominate many existing evaluations, real-world Vietnamese text is considerably less standardized. As a result, it remains unclear which retrieval methods are most effective for Vietnamese in practice. Existing studies typically evaluate on clean datasets or narrow domains (Ha et al., 2024; Pham Duy and Le Thanh, 2023; Nguyen et al., 2024), producing fragmented findings and offering little holistic understanding of Vietnamese retrieval effectiveness.

This work addresses that gap by presenting the first comprehensive benchmark of retrieval methods for Vietnamese across multiple domains. Our study is guided by a central research question:

Which retrieval methods are most effective for Vietnamese in realistic, multi-domain settings? Our contributions can be summarized as follows.

- We conduct a systematic comparison of three families of retrieval methods: lexical, dense, and hybrid. The evaluation spans diverse domains — including education, legal, healthcare, customer support, lifestyle reviews, and a cross-domain corpus from Wikipedia — all selected to reflect realistic Vietnamese data sources and practical usage scenarios.
- We introduce two new datasets constructed from authentic Vietnamese text. The first focuses on educational counseling in the context of university admissions, and the second captures customer service interactions. Both datasets exhibit real-world linguistic variation and noise, enabling more faithful evaluations than prior curated benchmarks.
- We provide empirical insights into the effectiveness of lexical, dense, and hybrid retrieval methods across domains. These insights clarify trade-offs, highlight domain-specific challenges, and offer practical guidance for building robust Vietnamese QA systems.

Together, these contributions offer the first large-scale, practice-oriented perspective on retrieval for Vietnamese. Beyond benchmarking existing methods across multiple domains, we also release new datasets and distilled insights that can inform academic research, guide system designers, and support the practical deployment of RAG systems in low-resource languages.

2 Related Work

Research on information retrieval for Vietnamese is still limited and fragmented. Existing studies fall into two main lines: evaluations of retrieval methods within specific domains, and the development of Vietnamese QA corpora.

Domain-specific retrieval evaluations. Several works have examined methods such as TF–IDF, BM25, or dense embeddings in the legal domain (Ba et al., 2024; Ha et al., 2024; Pham Duy and Le Thanh, 2023; Khang et al., 2024). These efforts typically focus on building RAG-based legal QA systems or refining lexical retrievers within that narrow setting. More recent studies investigate dense embeddings for Vietnamese retrieval (T. and T., 2024; Nguyen et al., 2024, 2025), yet they remain confined to a single domain and lack comparative evaluations across different classes of methods. Consequently, their findings provide useful but domain-bound insights with limited generalizability.

Vietnamese QA corpora. In parallel, several large QA datasets have been released, including cross-domain Wikipedia-style corpora (Van Nguyen et al., 2020, 2022; Tran et al., 2024) and community-based datasets in areas such as healthcare. These corpora have mainly been used to evaluate machine reading models or fine-tuned LLMs, with little consideration of retrieval effectiveness. Moreover, their texts are typically clean and well-structured, lacking the informal queries, linguistic variation, and noise characteristic of real-world Vietnamese data such as customer service or educational counseling. As a result, they provide limited insights for studying retrieval under realistic conditions.

To our knowledge, there has been no benchmark that systematically compares lexical, dense, and hybrid retrieval methods for Vietnamese across diverse and realistic domains. Prior work either focuses on a single domain, most often law, or uses datasets that do not capture authentic usage scenarios. Our work fills this gap by presenting the first systematic large-scale benchmark of Vietnamese retrieval methods, supported by two new datasets that capture authentic counseling and customer service interactions.

3 Methodology

3.1 Formulation of RAG

RAG consists of two components: a retriever and a generator, as illustrated in Figure 1. Given a query q and a knowledge base represented as a document collection $\mathcal{D} = \{d_i\}_{i=1}^{N}$, the retriever ranks documents using a scoring function $s(q, d)$ and returns the top-k results, formalized in Equation (1).

$$\mathcal{S}_k \;=\; \mathrm{TopK}_{d\in\mathcal{D}}\; s(q,d). \qquad (1)$$

Let L_θ denote the generator, typically instantiated as an LLM. Conditioned on both the query and the retrieved documents $\mathcal{S}_k$, the generator produces an output sequence by maximizing the conditional likelihood in Equation (2).

$$y^* \;=\; \arg\max_{y}\; \log p_\theta(y \mid q, \mathcal{S}_k)\,. \qquad (2)$$

When no retrieval is performed ($k = 0$), the formulation reduces to a standard conditional language model in Equation (3).

$$y^* \;=\; \arg\max_{y} \; \log p_\theta(y \mid q)\,. \tag{3}$$

Equation (2) and Equation (3) illustrate that retrieval is essential: without external documents, RAG collapses into a plain LLM. Meanwhile, Equation (1) shows that retrieval reduces to ranking documents according to $s(q, d)$, with different paradigms corresponding to different instantiations of s (lexical, dense, or hybrid). This choice is particularly critical for Vietnamese, where queries are often informal and documents noisy.

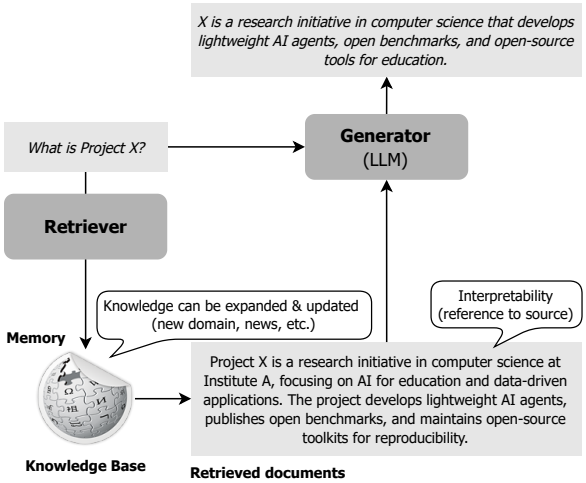

Figure 1: Overview of RAG. The retriever fetches relevant documents from a knowledge base, which are then provided to the generator (LLM) as additional context.

3.2 Benchmarking Setup

Benchmarking retrieval methods can be formalized as the task of identifying the scoring function $s(q, d)$ that maximizes retrieval performance across benchmark datasets. Formally, a benchmark is represented as in Equation (4).

$$\mathcal{B} = \{(q_j, RD_j, \mathcal{D})\}_{j=1}^{|\mathcal{B}|}, \tag{4}$$

where q_j is a query, RD_j is the set of ground-truth relevant documents, and $\mathcal{D}$ is the full document corpus shared across all queries. Extending the retrieval formulation in Equation (1), for each query j the retriever induced by s returns

$$\mathcal{S}_k^j(s) = \mathrm{TopK}_{d \in \mathcal{D}}\, s(q_j, d). \tag{5}$$

Let $\mathcal{M}(\cdot)$ denote an evaluation metric. The average benchmark score of s across $\mathcal{B}$ is defined in Equation (6).

$$\mathrm{Score}(s) = \frac{1}{|\mathcal{B}|} \sum_{j=1}^{|\mathcal{B}|} \mathcal{M}\big(q_j, RD_j, \mathcal{S}_k^j(s)\big). \tag{6}$$

The final objective is to identify the scoring function that maximizes this score, as given in Equation (7).

$$s^* = \arg\max_{s \in \mathcal{S}} \; \mathrm{Score}(s), \tag{7}$$

where $\mathcal{S}$ denotes the set of candidate retrieval methods (e.g., lexical, dense, hybrid). This formalization provides a unified framework for rigorous and fair comparison of retrieval paradigms under realistic conditions.

3.3 Retrieval Methods

Equation (1) formalizes retrieval as ranking documents according to a relevance function $s(q, d)$. As established in Equation (7), benchmarking seeks the scoring function s^* that maximizes retrieval effectiveness. In practice, retrieval methods can be broadly categorized into two families: sparse and dense models (Fan et al., 2024).

3.3.1 Lexical-based Retrieval

Lexical methods map each text into a sparse vector over a fixed vocabulary V, as defined in Equation (8).

$$f : \; \mathbb{T} \;\mapsto\; \mathbf{v} \in \mathbb{R}^{|V|}, \tag{8}$$

where $\mathbb{T}$ denotes the textual space and $\mathbf{v}$ is a bag-of-words vector indexed by terms in V. Given two representations $f(q)$ and $f(d)$, their similarity is typically computed by cosine similarity, as shown in Equation (9).

$$s(q, d) \;=\; \frac{f(q)^\top f(d)}{\|f(q)\|_2 \, \|f(d)\|_2}. \tag{9}$$

Two widely adopted instantiations of $f(\cdot)$ are TF–IDF (Salton and Buckley, 1988) and BM25 (Robertson and Zaragoza, 2009).

TF–IDF. Each document d is represented as a weighted vector $f(d) = (w_t(d))_{t \in V}$, where the weight of term t is given by Equation (10).

$$w_t^{\mathrm{tfidf}}(d) \;=\; \mathrm{tf}(t, d) \cdot \log \frac{N}{\mathrm{df}(t)}, \tag{10}$$

with $\mathrm{tf}(t, d)$ the frequency of t in d, $\mathrm{df}(t)$ the number of documents containing t, and N the corpus

size. Substituting these weights into the cosine similarity of Equation (9) yields the TF–IDF scoring function, as shown in Equation (11).

$$s_{\text{tf-idf}}(q,d) \;=\; \frac{\sum_{t\in q\cap d} w_t^{\text{tfidf}}(q)\, w_t^{\text{tfidf}}(d)}{\|f(q)\|_2\, \|f(d)\|_2}. \quad (11)$$

BM25. BM25 extends TF–IDF by introducing term-frequency saturation and document-length normalization. The weight of term t in document d is specified in Equation (12).

$$w_t^{\text{bm25}}(d) = \log \frac{N - \text{df}(t) + 0.5}{\text{df}(t) + 0.5} \times \frac{\text{tf}(t,d)(k_1+1)}{\text{tf}(t,d) + k_1\left(1 - b + b\,\frac{|d|}{\overline{L}}\right)}, \quad (12)$$

where $|d|$ is the document length, $\overline{L}$ the average length, and typical hyperparameters are $k_1 \in [1.2, 2.0]$ and $b \approx 0.75$. Unlike TF–IDF, BM25 does not rely on cosine normalization; instead, the overall score is directly defined as in Equation (13).

$$s_{\text{bm25}}(q,d) \;=\; \sum_{t\in q\cap d} w_t^{\text{bm25}}(d). \quad (13)$$

A key advantage of lexical methods is their simplicity and universality: once appropriate tokenization is available, they can be directly applied to Vietnamese retrieval tasks without training, relying solely on corpus-level statistics.

3.3.2 Dense-based Retrieval

Unlike lexical approaches that rely on exact term overlap, dense methods employ neural encoders to capture semantic similarity. Formally, an encoder $E(\cdot)$ maps text into an m-dimensional vector space, as shown in Equation 14.

$$E:\ \mathbb{T} \;\mapsto\; \mathbf{v} \in \mathbb{R}^m, \quad (14)$$

where m denotes the embedding dimension. Given query and document embeddings, relevance is typically measured by cosine similarity

$$s_{\text{dense}}(q,d) \;=\; \frac{E(q)^\top E(d)}{\|E(q)\|_2\, \|E(d)\|_2}. \quad (15)$$

Dense retrievers are commonly instantiated by bi-encoder architectures such as *Dense Passage Retrieval* (DPR) (Karpukhin et al., 2020), where queries and documents are encoded independently by $E(\cdot)$. Their training usually employs contrastive objectives, which encourage higher similarity for query–positive pairs than for query–negative pairs. A widely adopted formulation is the InfoNCE loss, given in Equation 16.

$$\mathcal{L}_{\text{NCE}} = -\log \frac{e^{\,s_{\text{dense}}(q,d^+)}}{e^{\,s_{\text{dense}}(q,d^+)} + \sum_{d^-\in\mathcal{N}} e^{\,s_{\text{dense}}(q,d^-)}}, \quad (16)$$

where d^+ denotes the ground-truth relevant document and $\mathcal{N}$ is a set of negatives.

For Vietnamese retrieval, dense models can be initialized from multilingual encoders such as mBERT (Pires et al., 2019), XLM-R (Conneau et al., 2020), or BGE-M3 (Chen et al., 2024), as well as Vietnamese-specific encoders like PhoBERT (Nguyen and Tuan Nguyen, 2020), then fine-tuned with contrastive learning on task-specific data. Although more resource-intensive than sparse methods, dense retrieval yields stronger semantic generalization, particularly across paraphrases and morphologically rich expressions.

3.4 Hybrid Retrieval

Hybrid methods aim to combine the complementary strengths of lexical and dense retrieval. Two widely used strategies are weighted fusion and rank-based fusion.

Weighted Fusion. Scores from lexical and dense retrievers can be linearly combined, as shown in Equation (17).

$$s_{\text{hybrid}}(q,d) = \alpha\, s_{\text{lexical}}(q,d) + (1-\alpha)\, s_{\text{dense}}(q,d), \quad (17)$$

where $\alpha \in [0,1]$ is a tunable weight. Here, s_{lexical} may correspond to s_{tfidf} or s_{bm25}. Since these scores are not directly comparable in scale, normalization (e.g., min–max or z-score) is typically applied before fusion.

Rank-based Fusion. An alternative is to fuse ranks rather than scores. A widely used method is *Reciprocal Rank Fusion* (RRF) (Cormack et al., 2009), which assigns each document d the fused score in Equation (18).

$$s_{\text{RRF}}(q,d) = \sum_{m\in\{\text{lexical},\text{dense}\}} \frac{1}{c + \text{rank}_m(q,d)}, \quad (18)$$

where $\text{rank}_m(q,d)$ is the rank of d under method m, and c is a smoothing constant (commonly $c = 60$). RRF is simple, parameter-free, and robust across heterogeneous retrieval models.

In practice, for Vietnamese retrieval, hybrid methods often outperform either lexical or dense

approaches alone. They simultaneously benefit from the precision of lexical matching and the semantic coverage of dense models, which is particularly advantageous in domain-specific applications such as legal and educational QA systems.

4 Experimentations

We design a comprehensive benchmark to evaluate retrieval methods for Vietnamese under realistic conditions. The study covers lexical, dense, and hybrid retrievers across multiple domains, and the results are analyzed to distill empirical insights that clarify trade-offs, highlight domain-specific challenges, and offer practical guidance for building robust Vietnamese QA systems.

4.1 Evaluation Metrics

Retrieval effectiveness is assessed using three standard metrics: Precision@k, Recall@k, and *Mean Reciprocal Rank* (MRR) (Yu et al., 2025).

Precision@k. This metric evaluates accuracy by measuring the proportion of retrieved documents that are relevant. Formally, it is defined in Equation (19).

$$\text{Precision@}k(q_j) \;=\; \frac{|\,\mathcal{S}_k^j(s) \cap RD_j\,|}{k}, \qquad (19)$$

Recall@k. Recall@k quantifies coverage by computing the fraction of relevant documents retrieved within the top-k, as in Equation (20).

$$\text{Recall@}k(q_j) \;=\; \frac{|\,\mathcal{S}_k^j(s) \cap RD_j\,|}{|RD_j|}. \qquad (20)$$

MRR@k. Finally, MRR@k reflects ranking fidelity by rewarding systems that return relevant documents earlier. Let rank_j denote the position of the first relevant document for query q_j. Equation (21) gives the formal definition:

$$\text{MRR@}k \;=\; \frac{1}{|Q|}\sum_{j=1}^{|Q|} \frac{\mathbf{1}[\text{rank}_j \le k]}{\text{rank}_j}, \qquad (21)$$

where $|Q|$ is the number of queries and $\mathbf{1}[\cdot]$ is the indicator function, equal to 1 if the condition holds and 0 otherwise.

4.2 Datasets

We curated a benchmark where each dataset reflects naturally occurring queries paired with domain-specific documents, ensuring both linguistic diversity and retrieval difficulty.

Education. We compiled authentic questions posed by university students on topics such as admissions, academic regulations, and campus policies. These were aligned with institutional documents segmented into coherent chunks, forming the *Educational Counseling QA* (EduCoQA) dataset. We also incorporated *ViRHE4QA* (Do et al., 2025), a public dataset on higher education rules.

Customer Support. To capture the nuances of human interaction, we collected queries from real-world exchanges between customers and service agents. Each query was linked to relevant materials such as brochures, policy manuals, and troubleshooting guides, resulting in the *Customer Support Conversations Dataset* (CSConDa).

Legal. We adopted two established Vietnamese legal retrieval benchmarks: (i) the *Automated Legal Question Answering Competition* (ALQAC)[1], and (ii) the *Zalo Legal Text Retrieval Challenge*[2]. Together, they provide a representative testbed for statutory and regulatory document retrieval.

Healthcare. We evaluated two medical QA datasets: (i) *ViNewsQA* (Van Nguyen et al., 2022), derived from Vietnamese healthcare news articles, and (ii) *ViMedAQA* (Tran et al., 2024), spanning four subtopics—body parts, diseases, drugs, and treatments. Together, they cover both consumer and professional healthcare needs.

Lifestyle and Reviews. We included two datasets capturing informal, everyday queries: (i) *VlogQA* (Ngo et al., 2024), based on transcripts of Vietnamese lifestyle vlogs, and (ii) *ViRe4MRC* (Do et al., 2023), derived from food and technology product reviews. Both highlight naturally phrased queries in non-technical contexts.

Cross-domain Open Knowledge. As a broad-coverage benchmark, we used *UIT-ViQuAD* (Van Nguyen et al., 2020), a large-scale Vietnamese QA dataset constructed from Wikipedia articles. This dataset complements the domain-specific corpora by providing open-domain queries.

Sampling Strategy. Each dataset was standardized to 1000 query–document pairs. We removed duplicate contexts, sampled one query per unique context to maximize diversity, and filled any short-

[1] https://alqac.github.io
[2] https://challenge.zalo.ai/portal/legal-text-retrieval

fall from the residual pool. Gold relevance labels were then remapped to the deduplicated corpus.

4.3 Baselines

We benchmark three families of retrieval methods introduced in Section 3.3. The selection covers both proprietary and open-source models, ensuring a balanced comparison across resource conditions.

Lexical. We adopt TF–IDF and BM25 as canonical sparse baselines.

Dense. We evaluate models in three categories: proprietary, multilingual, and Vietnamese-specific.

(i) Commercial. text-embedding-3-large, OpenAI's latest embedding model optimized for multilingual tasks, serves as the proprietary baseline.

(ii) Open-source Multilingual. We include BAAI/bge-m3 (Chen et al., 2024), a state-of-the-art model covering over 100 languages and excelling at cross-lingual and long-document retrieval. We also add paraphrase-multilingual-MiniLM-L12-v2, a compact model trained under the Sentence-BERT framework (Reimers and Gurevych, 2019) for sentence-level similarity, and despite its smaller size it remains one of the most widely used multilingual embeddings, including for Vietnamese.[3]

(iii) Vietnamese-specific. To capture linguistic phenomena unique to Vietnamese, we evaluate three models that reflect the progression of local embedding research. bkai-foundation-models/vietnamese-bi-encoder extends *PhoBERT* (Nguyen and Tuan Nguyen, 2020), the first large-scale monolingual transformer for Vietnamese pre-trained on a 20GB news corpus, into a bi-encoder for retrieval; dangvantuan/vietnamese-document-embedding adapts *mGTE* (Zhang et al., 2024), a multilingual generative embedding model, to Vietnamese long documents through multi-stage fine-tuning; and AITeamVN/Vietnamese_Embedding_v2, the latest community model, fine-tuned from bge-m3 on 1.1M query–document pairs with hard negatives, is currently the strongest available Vietnamese embedding model for retrieval.

Hybrid. Hybrid retrievers integrate sparse and dense signals via two strategies: (i) weighted score interpolation and (ii) rank-based fusion.

[3] https://huggingface.co/models?pipeline_tag=sentence-similarity&language=vi&sort=trending

4.4 Results and Analysis

We evaluate sparse, dense, and hybrid retrievers across six domains. Table 1, Table 2, and Table 3 present the consolidated results, capturing both domain-specific behaviors and consistent performance trends. In this section, we distill the key insights that emerge from these comparisons, while detailed numerical breakdowns and per-domain analyses are provided in Appendix E.

I1. Commercial multilingual embeddings fall short of Vietnamese-specific encoders, highlighting the importance of language-focused training.

I2. Hybrids that integrate BM25 with dense encoders consistently outperform both dense-only and sparse-only baselines, confirming the value of combining lexical and semantic signals.

I3. Among Vietnamese-specific encoders, Vietnamese_Embedding_v2 paired with BM25 delivers the strongest overall results, setting a new benchmark for local retrieval models.

I4. Informal and user-generated domains remain the most difficult: even the best hybrids achieve only modest accuracy, underscoring the challenges of paraphrasing, colloquial vocabulary, and noisy contexts.

I5. In structured corpora such as Legal and Wikipedia, lexical signals remain indispensable. BM25 already provides a strong baseline, and hybrids push performance close to the ceiling.

I6. In conversational or paraphrastic settings, multilingual encoders can rival—or occasionally surpass—Vietnamese-specific models, likely due to their broader cross-lingual exposure.

I7. Retrieval effectiveness is highly domain-sensitive: while structured datasets reach near-perfect scores, lifestyle and customer-support corpora expose persistent gaps.

I8. Dense-only encoders still hold value when lexical overlap is weak, offering robustness to semantic variation even without hybridization.

4.5 Error Analysis

We analyze representative failure cases from the weakest splits by absolute scores, including *EduCoQA* (Education), *CSConDa* (Customer Service), and *VlogQA*/*ViRe4MRC* (Lifestyle & Reviews). The errors highlight recurring linguistic and task-specific challenges that limit retrieval effectiveness. Detailed examples are provided in Appendix F.

Table 1: Retrieval results on Education (EduCoQA, ViRHE4QA) and Customer Service (CSConDa). Best scores are in **bold**, second-best underlined.

Domain	**Education**								**Customer Service**			
Method	EduCoQA				ViRHE4QA				CSConDa			
	P@1	R@10	R@20	MRR@10	P@1	R@10	R@20	MRR@10	P@1	R@10	R@20	MRR@10
TF–IDF	14.68	42.47	53.82	22.23	55.60	92.00	95.90	67.70	15.70	38.50	47.20	22.49
BM25	14.68	43.44	53.42	23.09	65.80	93.50	96.90	76.05	17.40	36.80	45.90	22.99
text-embedding-3-large												
Dense	20.16	51.86	63.01	30.30	52.60	88.80	93.60	64.94	33.70	56.80	63.80	41.06
+ BM25 (α)	22.11	57.34	65.95	32.54	66.70	95.40	97.90	76.74	**36.40**	60.20	66.20	**43.60**
+ TF–IDF (α)	22.70	57.34	66.54	32.40	62.90	94.00	97.40	73.97	34.90	**60.40**	**66.50**	42.45
BAAI/bge-m3												
Dense	**24.66**	55.77	64.77	**34.22**	59.20	91.10	95.10	70.40	30.80	53.90	61.00	37.98
+ BM25 (α)	22.90	**58.71**	**67.51**	33.68	71.10	95.90	98.20	79.94	33.90	56.90	63.90	40.97
+ TF–IDF (α)	23.68	57.93	**67.51**	33.78	65.90	95.30	97.40	76.55	33.10	57.00	63.80	40.67
+ TF–IDF (RRF)	20.55	53.82	67.32	29.72	61.80	94.60	97.70	73.72	28.40	54.20	63.90	35.82
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2												
Dense	14.48	40.70	52.05	20.90	30.80	63.00	71.90	40.43	11.80	30.00	39.10	16.71
bkai-foundation-models/vietnamese-bi-encoder												
Dense	18.79	48.53	58.51	27.01	46.80	77.80	85.50	56.58	15.70	34.90	41.70	21.09
dangvantuan/vietnamese-document-embedding												
Dense	20.55	53.42	63.21	30.76	50.80	86.70	92.20	63.01	28.40	53.00	59.90	36.12
AITeamVN/Vietnamese_Embedding_v2												
Dense	19.77	50.29	59.88	28.99	61.40	92.20	95.60	72.04	31.40	54.00	61.40	38.40
+ BM25 (α)	19.57	56.16	65.17	29.93	**72.50**	**96.90**	**98.80**	**81.42**	33.70	57.90	64.30	41.22
+ BM25 (RRF)	21.14	53.23	64.58	29.90	68.60	96.20	98.30	78.72	28.80	54.70	62.40	36.70

Table 2: Retrieval results on Legal (ALQAC, ZaloLegalQA) and Healthcare (ViNewsQA, ViMedAQA).

Domain	**Legal**								**Healthcare**							
Method	ALQAC				ZaloLegalQA				ViNewsQA				ViMedAQA			
	P@1	R@10	R@20	MRR@10	P@1	R@10	R@20	MRR@10	P@1	R@10	R@20	MRR@10	P@1	R@10	R@20	MRR@10
TF–IDF	82.83	96.23	98.11	88.34	64.70	92.47	96.45	75.07	52.20	79.10	84.70	60.93	61.50	84.80	88.20	69.46
BM25	89.25	97.92	99.25	92.20	71.40	92.18	94.42	79.39	59.00	80.20	84.30	66.13	65.40	84.50	87.30	71.36
text-embedding-3-large																
Dense	84.53	98.68	99.81	90.13	80.50	96.83	98.53	87.10	49.20	76.40	81.40	58.59	80.40	95.40	97.20	85.44
+ BM25 (α)	93.02	99.43	**100.00**	95.79	84.00	97.93	98.92	89.83	64.70	85.90	**90.80**	71.83	**83.00**	96.10	**97.90**	**87.37**
+ TF–IDF (α)	88.87	99.25	99.81	93.22	82.20	98.07	**99.27**	88.52	62.40	84.50	89.60	69.96	80.90	**96.30**	97.80	86.22
+ BM25 (RRF)	90.57	98.87	99.81	94.06	78.10	96.28	98.12	85.50	61.00	85.30	90.20	68.74	77.70	92.10	95.00	82.16
BAAI/bge-m3																
Dense	90.38	99.43	**100.00**	94.14	82.30	97.83	98.77	88.52	57.60	79.00	83.40	64.72	81.20	94.10	96.70	85.60
+ BM25 (α)	**94.72**	**99.62**	99.81	**96.66**	82.10	98.02	98.62	88.62	65.50	87.10	90.10	72.87	82.50	94.60	96.70	86.58
+ TF–IDF (α)	92.08	99.43	99.81	95.02	80.50	98.12	99.05	87.69	63.90	86.40	89.50	71.41	81.30	94.40	96.60	85.95
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2																
Dense	65.85	92.45	95.09	74.37	51.10	80.23	85.87	60.75	25.50	52.20	60.00	33.54	48.00	73.50	80.70	56.19
bkai-foundation-models/vietnamese-bi-encoder																
Dense	80.75	95.66	98.11	86.21	71.00	92.72	94.72	79.20	45.50	67.50	73.20	52.39	70.10	87.20	90.40	75.68
dangvantuan/vietnamese-document-embedding																
Dense	85.85	98.49	99.25	90.51	77.70	96.27	97.95	85.08	54.90	76.80	80.50	61.77	75.50	90.60	93.00	80.86
+ TF–IDF (α)	89.06	98.87	99.62	92.96	79.80	98.35	99.00	86.80	61.70	84.30	88.30	69.22	78.30	93.10	94.70	83.73
AITeamVN/Vietnamese_Embedding_v2																
Dense	90.38	99.06	99.81	93.96	86.20	98.32	98.97	91.04	55.40	78.20	82.90	63.01	76.90	93.50	96.10	82.40
+ BM25 (α)	93.77	99.25	**100.00**	95.95	**87.40**	98.17	98.67	**91.67**	**68.60**	**87.30**	90.00	**74.96**	81.40	93.80	95.70	85.38
+ TF–IDF (α)	92.64	**99.62**	99.81	95.26	85.40	**98.42**	99.10	90.91	65.30	86.10	90.00	72.35	79.80	94.10	96.20	84.61
+ BM25 (RRF)	92.08	98.68	99.81	94.78	81.00	96.27	97.75	87.12	65.10	86.30	89.90	72.00	76.60	90.80	94.40	81.43
+ TF–IDF (RRF)	90.19	98.49	99.81	93.15	79.10	97.65	98.60	86.65	62.10	85.40	89.90	69.71	75.00	91.80	94.60	80.92

E1. Colloquial and noisy user queries remain a major obstacle. Abbreviations (*"ptn" = phòng thí nghiệm [laboratory]*, *"khmt" = khoa học máy tính [computer science]*), slang (*"cx" = cũng [also]*, *"đc" = được [can/be possible]*, *"ak" = vậy à [oh really?]*), emojis (😊, 😡), filler tokens from transcripts (*"à à ừ ừ" [uh, um]*), and code-switching all reduce lexical overlap for sparse retrievers and introduce semantic drift for dense encoders. Spelling variation and typos and missing diacritics further degrade retrieval quality, while simple casing inconsistencies have similar negative effects.

E2. Style and register mismatches frequently mislead retrievers. Informal queries are often aligned with highly formal gold passages (for example, policy templates in education or customer service), while formal questions may be paired with noisy, colloquial answers (for example, product reviews). Such mismatches yield passages that are topically related but pragmatically irrelevant.

E3. Entity, scope, and intent ambiguity remains unresolved, especially in the education and

Table 3: Retrieval results on Lifestyle & Reviews (VlogQA, ViRe4MRC) and Cross-domain Open Knowledge (UIT-ViQuAD).

Domain	**Lifestyle & Reviews**								**Cross-domain Open Knowledge**			
Method	VlogQA				ViRe4MRC				UIT-ViQuAD			
	P@1	R@10	R@20	MRR@10	P@1	R@10	R@20	MRR@10	P@1	R@10	R@20	MRR@10
TF–IDF	13.40	34.60	47.00	19.55	3.70	17.50	23.70	7.24	50.00	91.00	94.00	64.57
BM25	18.00	39.50	45.50	23.90	6.60	20.40	26.70	10.25	70.80	91.60	93.90	78.09
text-embedding-3-large												
Dense	13.50	37.00	45.60	20.45	9.70	30.00	38.60	15.03	71.20	92.40	96.00	78.75
BAAI/bge-m3												
Dense	24.20	51.90	59.90	32.49	12.40	30.80	40.00	17.25	80.60	96.40	98.40	86.62
+ BM25 (α)	**30.50**	**59.20**	66.10	**39.20**	12.40	**31.10**	**40.80**	18.02	88.30	99.00	**99.30**	92.41
+ TF–IDF (α)	27.20	57.70	64.90	36.19	12.00	30.90	39.50	17.64	83.20	98.70	99.20	89.48
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2												
Dense	4.50	14.80	20.50	7.16	4.70	16.10	22.10	7.30	55.90	81.30	87.70	64.15
bkai-foundation-models/vietnamese-bi-encoder												
Dense	13.90	34.30	42.80	19.46	8.60	21.10	29.00	11.97	68.00	88.40	92.10	74.62
dangvantuan/vietnamese-document-embedding												
Dense	22.70	46.90	56.20	29.62	10.70	27.40	35.70	15.02	75.70	95.60	97.40	83.00
+ BM25 (α)	28.80	57.50	**66.90**	36.96	11.60	29.70	38.20	16.55	85.50	98.70	**99.30**	90.34
AITeamVN/Vietnamese_Embedding_v2												
Dense	22.30	49.00	57.50	29.92	10.60	28.60	38.70	15.48	82.20	98.10	99.00	88.24
+ BM25 (α)	29.20	57.80	65.60	37.80	**13.00**	30.80	39.30	**18.21**	**89.10**	99.20	**99.30**	**93.00**
+ TF–IDF (α)	25.80	55.70	63.40	35.04	12.20	30.60	40.00	17.40	84.40	**99.30**	**99.30**	90.18
+ BM25 (RRF)	27.10	56.40	65.20	35.80	**13.00**	29.40	38.30	17.56	82.10	97.50	99.10	87.95
+ TF–IDF (RRF)	24.10	55.20	63.20	33.67	12.20	29.40	37.10	16.98	79.00	97.60	**99.30**	86.02

customer-service domains. Underspecified queries such as *"trưởng khoa là ai?" [who is the dean?]* can point to multiple valid entities, while short paraphrastic one-liners (*"e gọi cho a đx k?" [can I call you?]*) provide too little signal. Heterogeneous intents in customer-service datasets (pricing, eligibility, compliance) exacerbate the risk of retrieving passages that match surface terms but fail to capture the user's actual information need.

5 Discussion

Our findings reveal several broader implications for Vietnamese QA and cross-lingual retrieval.

Implications for system design. The consistent gains of BM25 hybrids confirm that hybridization should be the default strategy for Vietnamese QA. Vietnamese-focused encoders excel in structured corpora, while multilingual encoders remain strong in conversational and customer-support datasets. Adaptive pipelines that select or combine retrievers by domain characteristics are therefore more effective than a uniform approach.

Limitations and open challenges. Vietnamese retrieval still faces clear gaps between structured and informal datasets. While Legal benchmarks reach near-ceiling recall, corpora such as VlogQA, ViRe4MRC, EduCoQA, and CSConDa plateau at much lower levels. These gaps reflect persistent difficulties highlighted in our error analysis, including colloquial and noisy queries, style mismatches, and underspecified or ambiguous intents. In addition, most Vietnamese-specific encoders remain lightweight compared to commercial models, and multimodal scenarios such as spoken queries and lecture transcripts, which are common in real-world applications, are not yet adequately addressed.

Future directions. Promising avenues include domain-adaptive fine-tuning, paraphrase-augmented training, and refining score interpolation hybrids, which proved more stable than rank-based fusion. Expanding to multimodal retrieval, particularly for education and healthcare, would better reflect real-world usage. Enlarging Vietnamese retrieval datasets with richer user-generated content is also essential for robustness and fairness.

6 Conclusion

This study introduced a large-scale benchmark of Vietnamese retrieval across six domains, comparing lexical, dense, and hybrid methods and releasing two datasets in education and customer service. Results show that BM25-based hybrids provide the most reliable backbone, Vietnamese-specific encoders excel in structured corpora, and multilingual models remain competitive in conversational settings, while informal user-generated data pose challenges. These findings provide practical guidance for building robust Vietnamese QA systems and point to future directions in domain-adaptive fine-tuning, paraphrase augmentation, multimodal retrieval, and dataset expansion to strengthen retrieval effectiveness in low-resource languages.

Limitations

Although our benchmark represents the most comprehensive evaluation of Vietnamese retrieval to date, several aspects remain open for future work. First, our focus is on text-based retrieval, while multimodal scenarios such as speech or video transcripts—highly relevant in education and healthcare—are not yet included. Second, current Vietnamese encoders are relatively compact compared to large commercial counterparts, which may affect scalability for very large corpora, though their efficiency makes them attractive for resource-constrained applications. Third, despite the addition of two new datasets, informal and noisy user-generated content is still underrepresented, highlighting an important area for further expansion. These limitations are natural given the scope of this work and point to promising directions for advancing robust and inclusive Vietnamese QA.

Supplementary Materials Availability Statement

All datasets, including both public and proposed ones, as well as evaluation subsets and benchmark scripts, are available at the anonymous repository: https://anonymous.4open.science/r/ViRE for review purposes. The repository will be made publicly accessible upon acceptance. All reused resources follow their original academic or open-source licenses. The proposed datasets (EduCoQA and CSConDa) will be released under the CC BY-NC 4.0 license for research use only.

References

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A Survey on RAG with LLMs. *Procedia Computer Science*, 246:3781–3790. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024).

Thiem Nguyen Ba, Vinh Doan The, Tung Pham Quang, and Toan Tran Van. 2024. Vietnamese Legal Information Retrieval in Question-Answering System.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.

Tinh Pham Phuc Do, Ngoc Dinh Duy Cao, Nhan Thanh Nguyen, Tin Van Huynh, and Kiet Van Nguyen. 2023. Machine Reading Comprehension for Vietnamese Customer Reviews: Task, Corpus and Baseline Models. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 24–35, Hong Kong, China. Association for Computational Linguistics.

Tinh Pham Phuc Do, Ngoc Dinh Duy Cao, Khanh Quoc Tran, and Kiet Van Nguyen. 2025. R2GQA: retriever-reader-generator question answering system to support students understanding legal regulations in higher education. *Artificial Intelligence and Law*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6491–6501, New York, NY, USA. Association for Computing Machinery.

Nguyen Thu Ha, Truong-Phuc Nguyen, Khang T. Trung, Huu-Loi Le, Le Thi Viet Huong, Chi Thanh Nguyen, and Minh-Tien Nguyen. 2024. Vietnamese Legal Question Answering: An Experimental Study. In *2024 16th International Conference on Knowledge and System Engineering (KSE)*, pages 440–446.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and

Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Nguyen Hoang Gia Khang, Nguyen Minh Nhat, Trung Nguyen Quoc, and Vinh Truong Hoang. 2024. Vietnamese Legal Text Retrieval based on Sparse and Dense Retrieval approaches. *Procedia Computer Science*, 234:196–203. Seventh Information Systems International Conference (ISICO 2023).

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards General Text Embeddings with Multi-stage Contrastive Learning.

Thinh Ngo, Khoa Dang, Son Luu, Kiet Nguyen, and Ngan Nguyen. 2024. VlogQA: Task, Dataset, and Baseline Models for Vietnamese Spoken-Based Machine Reading Comprehension. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1310–1324, St. Julian's, Malta. Association for Computational Linguistics.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Long S. T. Nguyen and Tho T. Quan. 2025. URAG: Implementing a Unified Hybrid RAG for Precise Answers in University Admission Chatbots – A Case Study at HCMUT. In *Information and Communication Technology*, pages 82–93, Singapore. Springer Nature Singapore.

Toan Ngoc Nguyen, Nam Le Hai, Nguyen Doan Hieu, Dai An Nguyen, Linh Ngo Van, Thien Huu Nguyen, and Sang Dinh. 2025. Improving Vietnamese-English Cross-Lingual Retrieval for Legal and General Domains. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 142–153, Albuquerque, New Mexico. Association for Computational Linguistics.

Vinh Nguyen, Nam Tran, Long Nguyen, and Dien Dinh. 2024. Advancing Vietnamese Information Retrieval with Learning Objective and Benchmark. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 46–56, Tokyo, Japan. Tokyo University of Foreign Studies.

Anh Pham Duy and Huong Le Thanh. 2023. A Question-Answering System for Vietnamese Public Administrative Services. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, SOICT '23, page 85–92, New York, NY, USA. Association for Computing Machinery.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Hai Nguyen T. and Huong Le T. 2024. Enhancing ColBERT: A Method for Reducing Space Complexity and Accelerating Retrieval Speed. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 820–829, Tokyo, Japan. Tokyo University of Foreign Studies.

Minh-Nam Tran, Phu-Vinh Nguyen, Long Nguyen, and Dien Dinh. 2024. ViMedAQA: A Vietnamese Medical Abstractive Question-Answering Dataset and Findings of Large Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 252–260, Bangkok, Thailand. Association for Computational Linguistics.

Kiet Van Nguyen, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020. A Vietnamese Dataset for Evaluating Machine Reading Comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kiet Van Nguyen, Tin Van Huynh, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen.

2022. New Vietnamese Corpus for Machine Reading Comprehension of Health News Articles. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. Evaluation of Retrieval-Augmented Generation: A Survey. In *Big Data*, pages 102–120, Singapore. Springer Nature Singapore.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

A Experimental Setup

All experiments were conducted on an NVIDIA RTX 6000 GPU with 24GB VRAM. Open-source dense encoders were evaluated directly using their official Hugging Face checkpoints, while the proprietary OpenAI model was accessed via API. The key hyperparameter settings for BM25 and hybrid retrieval methods are summarized in Table 4.

Table 4: Hyperparameter settings for retrieval models.

Parameter	Value
BM25 k_1	1.5
BM25 b	0.75
Weighted fusion coefficient (α)	0.7
RRF constant (c)	60

B Model Analysis

We analyze the architectures and representational capacities of the embedding models employed in our experiments, as summarized in Table 5. The models span diverse architectural families and differ markedly in parameter scale and embedding dimensionality, enabling a systematic examination of how these factors influence retrieval effectiveness across domains.

Multilingual backbones such as XLM-RoBERTa (Conneau et al., 2020) and *General Text Embeddings* (GTE) (Li et al., 2023) demonstrate stronger cross-domain generalization by leveraging broader representational spaces (768–1024 dimensions) and richer semantic priors learned from cross-lingual corpora. Among them, Vietnamese_Embedding_v2, despite sharing a similar parameter scale with bge-m3, consistently delivers superior and more stable results, indicating that language-aligned pretraining contributes more to retrieval quality than sheer model size. Overall, these findings confirm that parameter scale and embedding dimensionality jointly govern retrieval effectiveness: medium-to-large encoders with 1024-dimensional embeddings provide the best balance between expressiveness, efficiency, and robustness for Vietnamese retrieval tasks.

C Data Analysis

We provide a detailed statistical analysis of the datasets used in our benchmark. Table 6 reports the number of records, unique contexts, and token-level statistics of queries and contexts. For tokenization, we adopt the Vietnamese tokenizer from the widely used Underthesea NLP toolkit[4], which provides reliable segmentation for Vietnamese text.

It is evident from Table 6 that both query and context lengths vary widely across datasets. For example, contexts in VlogQA average over 2,200 tokens, while those in ViRe4MRC average fewer than 100. In contrast, query lengths remain relatively short across domains, with means ranging from about 9 to 19 tokens. Importantly, retrieval effectiveness does not correlate directly with context length: datasets with very long contexts (e.g., VlogQA) remain challenging, while shorter but more structured corpora (e.g., Legal) achieve near-ceiling performance.

D Proposed Datasets

To enable a systematic evaluation of Vietnamese retrieval models, we introduce two new benchmarks: ***Edu**cational **Co**unseling **QA*** (EduCoQA) and ***C**ustomer **S**upport **Con**versations **Da**taset* (CSConDa). Both datasets are designed to capture realistic, domain-specific information needs while preserving the linguistic complexity of user-generated Vietnamese. Below we detail their construction and topical coverage.

D.1 CSConDa

Data Collection. CSConDa was created in collaboration with a leading national provider of multi-channel customer service solutions. The dataset was built through a three-phase pipeline:

[4] https://github.com/undertheseanlp/underthesea

Table 5: Embedding model specifications.

Model ID	Architecture	#Params	Embedding Dim.
OpenAI's text-embedding-3-large	–	–	3072
BAAI/bge-m3	XLM-RoBERTa	567,754,752	1024
paraphrase-multilingual-MiniLM-L12-v2	BERT	117,653,760	384
bkai-foundation-models/vietnamese-bi-encoder	RoBERTa	134,998,272	768
dangvantuan/vietnamese-document-embedding	GTE (Li et al., 2023)	305,368,320	768
AITeamVN/Vietnamese_Embedding_v2	XLM-RoBERTa	567,754,752	1024

Table 6: Dataset statistics (token-level).

Dataset	Records	Contexts	Query (min/mean/max)	Context (min/mean/max)
EduCoQA	511	262	3 / 11.68 / 46	5 / 144.97 / 513
ViRHE4QA	1000	297	4 / 14.12 / 46	13 / 268.48 / 1049
CSConDa	1000	1000	2 / 16.51 / 105	106 / 144.44 / 195
ALQAC	530	304	4 / 19.13 / 73	16 / 167.27 / 997
ZaloLegalQA	1000	1000	4 / 13.39 / 28	13 / 306.73 / 3310
ViNewsQA	1000	1000	4 / 10.41 / 27	90 / 334.76 / 694
ViMedAQA	1000	1000	4 / 11.47 / 36	10 / 97.97 / 547
VlogQA	1000	1000	4 / 9.99 / 22	216 / 2203.69 / 3807
ViRe4MRC	1000	1000	4 / 8.95 / 19	15 / 84.09 / 194
UIT-ViQuAD	1000	1000	2 / 11.75 / 25	74 / 147.93 / 604

- **Phase 1: Conversation harvesting.** Raw chat logs were collected from multi-channel support systems where customers interact with advisors across Facebook[5], Zalo[6], Shopee[7], and other platforms. Annotators applied strict criteria to retain coherent, high-quality conversations, while discarding sensitive or inappropriate content.

- **Phase 2: QA extraction.** An automated pipeline segmented conversations into contextually consistent QA pairs. Each pair was anonymized and carefully cleaned to remove personally identifiable information, emojis, and system-generated artifacts.

- **Phase 3: Document alignment.** Each question was linked to supporting passages drawn from brochures, help-center articles, and policy manuals. Segmentation ensured that reference documents contained precise spans sufficient to answer user queries.

[5]https://www.facebook.com/
[6]https://zalo.me/
[7]https://shopee.vn/

Topics. CSConDa spans a broad range of customer-support intents, including pricing inquiries, subscription eligibility, technical troubleshooting, account management, and policy compliance. The word cloud in Figure 2 highlights frequent keywords such as *"khách hàng"* (customer), *"quản lý"* (management), *"tin nhắn"* (message), *"hỗ trợ"* (support), and *"tài khoản"* (account), illustrating the dataset's coverage of practical service and operational contexts.

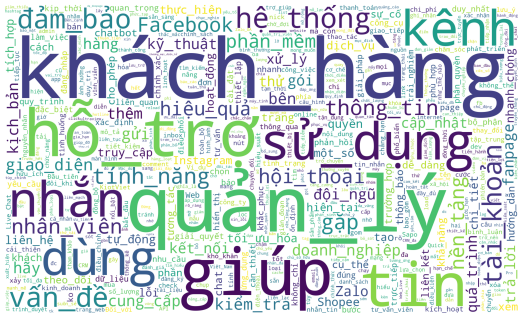

Figure 2: Word cloud of CSConDa, highlighting frequent customer-service topics.

D.2 EduCoQA

Data Collection. EduCoQA was curated from authentic university admission counseling sessions. To capture realistic student concerns, we deployed a lightweight RAG-based chatbot at educational fairs and institutional events. This setup enabled the collection of thousands of spontaneous questions from high-school students and parents. Data preparation proceeded in two phases:

- **Phase 1: Question collection.** Queries were logged during both in-person and online advising sessions. They reflect genuine concerns about admissions, academic programs, and institutional policies, providing a natural distribution of language and intent.

- **Phase 2: Document assignment.** Each query was aligned with authoritative sources, including official regulations, program descriptions, and departmental brochures. Answers provided by advisors were also incorporated. Segmentation ensured that retrieved passages were precise, verifiable, and grounded in institutional policy.

Topics. EduCoQA covers a broad range of counseling needs, from university and faculty information to admission criteria, scholarships, program structures, and career pathways. The word cloud in Figure 3 illustrates representative terms such as *"Kỹ thuật"* (Engineering), *"Bộ môn"* (Department), *"Chương trình"* (Program), *"Tuyển sinh"* (Admissions), and *"Sinh viên"* (Student).

Figure 3: Word cloud of EduCoQA, illustrating frequent topics in admission counseling.

D.3 Ethical Considerations

All data were anonymized and stripped of personally identifiable information. Customer-service logs were carefully filtered to exclude sensitive content, and all educational and policy documents are publicly accessible institutional resources. Both datasets are released exclusively for research purposes under terms that respect user privacy and institutional integrity. In addition, all annotators and advisors involved in dataset construction signed *Non-Disclosure Agreements* (NDAs) to ensure confidentiality and compliance with ethical standards.

D.4 Examples

Representative examples from both datasets are provided in Section F, which also serves as the failure analysis. These samples illustrate the linguistic variability, informality, and ambiguity that make EduCoQA and CSConDa challenging benchmarks for Vietnamese retrieval.

E Detail Analysis

We present detailed results and analyses for each domain, followed by an overall synthesis that grounds the insights reported in Section 4.4.

E.1 Education

Table 7 presents retrieval results on two datasets with contrasting characteristics: EduCoQA, which contains informal student counseling queries, and ViRHE4QA, which focuses on well-structured higher-education regulations.

First, dense encoders excel on informal counseling queries though overall scores remain modest. On EduCoQA, semantic models lead early-precision metrics: bge-m3 achieves P@1 = 24.66 and MRR@10 = 34.22, outperforming all hybrids, while text-embedding-3-large also performs competitively (20.16 / 30.30). Hybrids improve deeper recall (e.g., text-embedding-3-large + BM25 reaches R@50 = 80.63), but top-rank accuracy stays low across the board—highlighting the challenge of paraphrastic and colloquial counseling queries where lexical overlap is sparse. Second, lexical–dense hybrids dominate on structured regulations. On ViRHE4QA, Vietnamese Embedding v2 combined with BM25 (α) achieves the best overall balance (P@1 = 72.50, R@10 = 96.90, R@20 = 98.80, MRR@10 = 81.42). Multilingual bge-m3 + BM25 is also strong (71.10 / 79.94). Even the plain BM25 baseline remains competitive (65.80 / 76.05), underscoring the value of exact lexical matches in regulatory corpora.

Table 7: Retrieval results on **Education Domain** (EduCoQA and ViRHE4QA). Best in **bold**, second-best underlined.

Dataset	**EduCoQA**					**ViRHE4QA**				
Method	P@1	R@10	R@20	R@50	MRR@10	P@1	R@10	R@20	R@50	MRR@10
TF–IDF	14.68	42.47	53.82	67.12	22.23	55.60	92.00	95.90	99.10	67.70
BM25	14.68	43.44	53.42	66.93	23.09	65.80	93.50	96.90	98.70	76.05
text-embedding-3-large										
Dense	20.16	51.86	63.01	79.84	30.30	52.60	88.80	93.60	96.80	64.94
+ TF–IDF (α)	22.70	57.34	66.54	80.43	32.40	62.90	94.00	97.40	99.40	73.97
+ BM25 (α)	22.11	57.34	65.95	**80.63**	32.54	66.70	95.40	97.90	99.40	76.74
+ TF–IDF (RRF)	20.55	55.19	67.12	78.86	29.80	60.10	93.80	97.50	99.40	71.95
+ BM25 (RRF)	22.90	54.99	66.14	77.89	31.36	66.40	94.40	98.00	**99.70**	76.09
BAAI/bge-m3										
Dense	**24.66**	55.77	64.77	78.86	**34.22**	59.20	91.10	95.10	98.10	70.40
+ TF–IDF (α)	23.68	57.93	**67.51**	79.45	33.78	65.90	95.30	97.40	99.10	76.55
+ BM25 (α)	22.90	**58.71**	**67.51**	79.65	33.68	71.10	95.90	98.20	99.40	79.94
+ TF–IDF (RRF)	20.55	53.82	67.32	77.89	29.72	61.80	94.60	97.70	99.40	73.72
+ BM25 (RRF)	21.92	55.38	67.12	77.50	31.13	67.60	95.60	97.60	99.50	77.43
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2										
Dense	14.48	40.70	52.05	68.10	20.90	30.80	63.00	71.90	82.80	40.43
+ TF–IDF (α)	18.79	50.68	61.25	73.58	27.07	52.10	84.80	90.60	96.50	62.76
+ BM25 (α)	18.59	52.84	61.45	74.17	27.68	55.10	85.70	92.70	96.90	65.18
+ TF–IDF (RRF)	17.81	51.47	61.84	73.78	25.98	44.20	80.90	90.60	98.70	56.02
+ BM25 (RRF)	18.59	52.84	61.64	73.39	27.06	47.00	82.50	92.50	98.40	58.78
bkai-foundation-models/vietnamese-bi-encoder										
Dense	18.79	48.53	58.51	74.95	27.01	46.80	77.80	85.50	93.70	56.58
+ TF–IDF (α)	19.18	52.25	64.97	76.32	28.58	59.00	91.30	95.20	98.30	69.89
+ BM25 (α)	20.74	52.84	64.19	76.71	29.95	62.10	91.90	96.00	98.60	72.34
+ TF–IDF (RRF)	20.16	52.05	64.77	74.95	28.53	55.30	90.00	95.90	98.70	66.77
+ BM25 (RRF)	20.94	52.64	64.97	75.93	29.52	59.50	90.70	96.10	99.40	70.00
dangvantuan/vietnamese-document-embedding										
Dense	20.55	53.42	63.21	79.06	30.76	50.80	86.70	92.20	97.20	63.01
+ TF–IDF (α)	20.94	56.75	66.14	80.04	31.15	62.50	93.40	96.90	99.00	73.35
+ BM25 (α)	21.92	56.36	65.56	80.43	31.91	66.70	94.90	97.20	99.20	76.34
+ TF–IDF (RRF)	20.16	54.99	64.97	77.10	28.89	58.30	92.80	96.90	99.30	70.36
+ BM25 (RRF)	19.96	54.79	64.77	77.10	29.33	64.00	94.20	97.50	99.30	74.50
AITeamVN/Vietnamese_Embedding_v2										
Dense	19.77	50.29	59.88	75.15	28.99	61.40	92.20	95.60	98.20	72.04
+ TF–IDF (α)	18.98	56.36	65.36	79.45	29.33	68.90	96.00	97.70	99.10	78.93
+ BM25 (α)	19.57	56.16	65.17	79.26	29.93	**72.50**	**96.90**	**98.80**	99.30	**81.42**
+ TF–IDF (RRF)	19.37	53.82	65.36	77.69	28.80	63.80	95.80	97.80	99.30	75.41
+ BM25 (RRF)	21.14	53.23	64.58	77.89	29.90	68.60	96.20	98.30	99.50	78.72

E.2 Customer Service

Table 8 reports retrieval effectiveness on CSConDa, a customer–support corpus with short, paraphrastic queries.

First, commercial embeddings offer a clear advantage. As a dense-only retriever, OpenAI's model achieves P@1 = 33.70 and MRR@10 = 41.06, outperforming open-source models such as bge-m3 (30.80 / 37.98) and Vietnamese Embedding v2 (31.40 / 38.40). This reflects the benefit of large-scale conversational pretraining for intent-heavy, low-overlap queries in customer service. Second, hybridization with BM25 yields the strongest results. The best setup combines text-embedding-3-large with BM25 (interpolation), reaching P@1 = 36.40 and MRR@10 = 43.60. When fused with TF–IDF, the same model achieves the highest recall (R@10 = 60.40, R@20 = 66.50, R@50 = 75.90), confirming that lexical cues complement dense semantics. Open-source encoders

also gain from hybridization—e.g., Vietnamese Embedding v2 + BM25 attains MRR@10 = 41.22, and bge-m3 + BM25 reaches 40.97—yet both remain below the commercial baseline. Third, interpolation proves more stable than rank fusion. Across models, RRF underperforms α-weighted combinations: text-embedding-3-large + BM25 drops from 43.60 (interpolation) to 37.05 (RRF), with similar declines for bge-m3 (40.97 $\rightarrow$ 36.24) and Vietnamese Embedding v2 (41.22 $\rightarrow$ 36.70). Overall, while commercial embeddings and hybrids improve performance, absolute scores remain modest: even the best P@1 (36.40) and MRR@10 (43.60) stay below other domains. This underscores the difficulty of customer–support retrieval, where queries are short, paraphrastic, and offer limited lexical or semantic signal.

E.3 Legal

Table 9 reports retrieval effectiveness on ALQAC and ZaloLegalQA.

First, BM25 alone provides a very strong baseline. On ALQAC, it already achieves P@1 = 89.25 and MRR@10 = 92.20, underscoring the central role of lexical overlap in legal texts where terminology is highly standardized. Second, hybridization consistently improves performance. For example, bge-m3 combined with BM25 (α) reaches P@1 = 94.72 and MRR@10 = 96.66 on ALQAC, improving by more than +2.5 MRR points compared to its dense-only variant. On ZaloLegalQA, Vietnamese Embedding v2 paired with BM25 delivers the best balance (P@1 = 87.40, MRR@10 = 91.67), outperforming both sparse methods and multilingual encoders. Third, Vietnamese-specific encoders clearly dominate this domain. Both Vietnamese Embedding v2 and Dang Van Tuan's vietnamese-document-embedding, when fused with BM25, consistently surpass the OpenAI embedding (text-embedding-3-large) as well as multilingual baselines like bge-m3 and MiniLM. Finally, legal retrieval approaches a ceiling effect. Many systems achieve recall above 99%—sometimes hitting 100% at R@20 or R@50—showing that nearly all relevant passages can be retrieved. This makes the domain "easier" compared to others, but also less discriminative: improvements beyond BM25 hybrids are marginal.

E.4 Healthcare

Table 10 reports retrieval effectiveness on two datasets: ViNewsQA (news-style medical articles) and ViMedAQA (specialized clinical questions).

First, lexical methods provide strong baselines in news-style corpora. On ViNewsQA, BM25 already achieves solid effectiveness (P@1 = 59.00, MRR@10 = 66.13), reflecting substantial lexical overlap in journalistic text. Dense-only encoders underperform (e.g., text-embedding-3-large at 49.20 / 58.59), but hybrids clearly close the gap: Vietnamese_Embedding_v2 + BM25 reaches P@1 = 68.60 and MRR@10 = 74.96, the strongest configuration overall. Second, dense encoders are indispensable for specialized clinical QA. On ViMedAQA, semantic retrieval dominates: text-embedding-3-large alone achieves P@1 = 80.40 and MRR@10 = 85.44, outperforming all sparse baselines by more than +14 MRR points. Yet here too, hybridization pushes performance further: the same model combined with BM25 reaches P@1 = 83.00 and MRR@10 = 87.37, while vietnamese-document-embedding + BM25 attains 79.80 / 84.37. Third, interpolation consistently outperforms rank fusion. For example, bge-m3 + BM25 (α) achieves MRR@10 = 72.87 on ViNewsQA and 86.58 on ViMedAQA, compared to 70.63 and 81.59 with RRF. This indicates that weighted score fusion integrates lexical precision and semantic coverage more effectively than rank-based alternatives. Overall, the healthcare domain illustrates a twofold pattern: lexical methods remain highly competitive in news-style text, but hybrids are required to reach peak performance; meanwhile, dense encoders are essential for domain-specific clinical questions, with hybridization providing the final boost toward state-of-the-art results.

E.5 Lifestyles and Reviews

Table 11 presents retrieval results on two contrasting datasets: VlogQA (long-form multimodal vlogs) and ViRe4MRC (short review snippets).

First, dense encoders excel on long conversational queries. On VlogQA, bge-m3 achieves P@1 = 24.20 and MRR@10 = 32.49 in dense-only mode, outperforming both commercial and Vietnamese-specific encoders. Hybridization further boosts performance: bge-m3 + BM25 (α) reaches P@1 = 30.50 and MRR@10 = 39.20—the best overall. Vietnamese-specific embeddings such as AITeamVN/Vietnamese_Embedding_v2 (29.20 / 37.80) also perform competitively, suggesting that both multilingual pretraining and local adaptation are beneficial. Second, short review snippets yield low effectiveness. On ViRe4MRC, all models

Table 8: Retrieval results on **CSConDa**.

Method	P@1	R@10	R@20	R@50	MRR@10
TF–IDF	15.70	38.50	47.20	59.50	22.49
BM25	17.40	36.80	45.90	56.00	22.99
text-embedding-3-large					
Dense	33.70	56.80	63.80	73.10	41.06
+ TF–IDF (α)	34.90	**60.40**	**66.50**	**75.90**	42.45
+ BM25 (α)	**36.40**	60.20	66.20	74.70	**43.60**
+ TF–IDF (RRF)	28.80	55.00	63.80	74.30	36.45
+ BM25 (RRF)	29.60	54.40	64.00	73.00	37.05
BAAI/bge-m3					
Dense	30.80	53.90	61.00	69.90	37.98
+ TF–IDF (α)	33.10	57.00	63.80	73.10	40.67
+ BM25 (α)	33.90	56.90	63.90	72.80	40.97
+ TF–IDF (RRF)	28.40	54.20	63.90	71.60	35.82
+ BM25 (RRF)	28.60	53.70	62.80	71.10	36.24
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2					
Dense	11.80	30.00	39.10	49.50	16.71
+ TF–IDF (α)	19.60	45.30	51.40	62.50	27.05
+ BM25 (α)	18.90	45.40	51.60	61.70	26.18
+ TF–IDF (RRF)	17.70	43.80	54.30	64.90	25.03
+ BM25 (RRF)	17.50	43.80	53.30	64.80	24.63
bkai-foundation-models/vietnamese-bi-encoder					
Dense	15.70	34.90	41.70	53.40	21.09
+ TF–IDF (α)	22.20	46.00	52.40	63.00	28.95
+ BM25 (α)	22.70	45.70	52.80	62.50	28.93
+ TF–IDF (RRF)	19.10	44.20	55.00	65.80	26.34
+ BM25 (RRF)	20.00	44.90	54.30	65.30	26.98
dangvantuan/vietnamese-document-embedding					
Dense	28.40	53.00	59.90	68.30	36.12
+ TF–IDF (α)	31.10	57.90	65.50	73.80	39.46
+ BM25 (α)	32.40	57.80	64.60	73.10	40.05
+ TF–IDF (RRF)	26.00	51.80	64.00	74.60	33.88
+ BM25 (RRF)	26.70	52.60	63.70	74.40	34.65
AITeamVN/Vietnamese_Embedding_v2					
Dense	31.40	54.00	61.40	70.00	38.40
+ TF–IDF (α)	32.70	57.50	65.30	73.50	40.51
+ BM25 (α)	33.70	57.90	64.30	73.30	41.22
+ TF–IDF (RRF)	28.10	54.60	63.90	72.50	35.84
+ BM25 (RRF)	28.80	54.70	62.40	71.90	36.70

struggle: BM25 alone (P@1 = 6.60, MRR@10 = 10.25) performs nearly on par with dense-only retrieval, and even the strongest hybrid—Vietnamese Embedding v2 + BM25—only attains P@1 = 13.00 and MRR@10 = 18.21. These results are substantially lower than in other domains, indicating that fragmentary and noisy reviews offer limited lexical and semantic cues for retrieval. Overall, the lifestyle and reviews domain highlights an inherent difficulty: while dense and hybrid models improve over sparse baselines in long-form queries, short informal reviews remain challenging, with no method achieving high accuracy.

E.6 Cross-domain Open Knowledge

Table 12 reports results on UIT-ViQuAD, a dataset spanning diverse knowledge sources.

First, sparse methods remain remarkably competitive. BM25 alone reaches P@1 = 70.80 and MRR@10 = 78.09, outperforming multilingual MiniLM and rivaling several dense encoders. This indicates that exact lexical overlap continues to

Table 9: Retrieval results on **Legal Domain** (ALQAC and ZaloLegalQA).

Dataset	**ALQAC**					**ZaloLegalQA**				
Method	P@1	R@10	R@20	R@50	MRR@10	P@1	R@10	R@20	R@50	MRR@10
TF–IDF	82.83	96.23	98.11	99.25	88.34	64.70	92.47	96.45	97.70	75.07
BM25	89.25	97.92	99.25	99.62	92.20	71.40	92.18	94.42	96.30	79.39
text-embedding-3-large										
Dense	84.53	98.68	99.81	**100.00**	90.13	80.50	96.83	98.53	99.47	87.10
+ TF–IDF (α)	88.87	99.25	99.81	**100.00**	93.22	82.20	98.07	**99.27**	99.50	88.52
+ BM25 (α)	93.02	99.43	**100.00**	**100.00**	95.79	84.00	97.93	98.92	99.45	89.83
+ TF–IDF (RRF)	87.36	98.30	99.06	**100.00**	91.83	77.10	97.72	98.90	99.40	85.08
+ BM25 (RRF)	90.57	98.87	99.81	**100.00**	94.06	78.10	96.28	98.12	99.40	85.50
BAAI/bge-m3										
Dense	90.38	99.43	**100.00**	**100.00**	94.14	82.30	97.83	98.77	99.12	88.52
+ TF–IDF (α)	92.08	99.43	99.81	**100.00**	95.02	80.50	98.12	99.05	99.40	87.69
+ BM25 (α)	**94.72**	**99.62**	99.81	**100.00**	**96.66**	82.10	98.02	98.62	99.25	88.62
+ TF–IDF (RRF)	88.49	98.49	99.62	**100.00**	92.37	76.00	97.22	98.70	99.20	84.43
+ BM25 (RRF)	91.13	99.06	99.62	**100.00**	94.40	79.20	96.17	97.67	99.10	86.05
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2										
Dense	65.85	92.45	95.09	98.68	74.37	51.10	80.23	85.87	92.25	60.75
+ TF–IDF (α)	83.96	97.36	99.06	99.81	89.27	69.70	93.92	96.60	98.45	78.50
+ BM25 (α)	88.30	97.36	99.25	99.81	91.80	70.60	94.32	95.95	98.05	79.43
+ TF–IDF (RRF)	77.55	96.79	98.68	99.81	84.64	64.10	91.60	96.25	98.65	73.74
+ BM25 (RRF)	80.19	97.55	98.68	99.62	86.93	64.60	91.62	96.70	98.35	74.46
bkai-foundation-models/vietnamese-bi-encoder										
Dense	80.75	95.66	98.11	99.81	86.21	71.00	92.72	94.72	97.35	79.20
+ TF–IDF (α)	88.68	98.49	99.25	**100.00**	92.60	77.40	96.72	98.20	99.15	84.88
+ BM25 (α)	89.62	98.68	99.62	**100.00**	93.33	80.30	96.67	97.97	98.75	86.67
+ TF–IDF (RRF)	84.53	97.74	99.25	**100.00**	89.43	72.90	95.52	97.85	99.10	81.48
+ BM25 (RRF)	88.30	98.30	99.62	**100.00**	92.13	75.30	95.62	97.20	98.95	82.96
dangvantuan/vietnamese-document-embedding										
Dense	85.85	98.49	99.25	99.81	90.51	77.70	96.27	97.95	98.95	85.08
+ TF–IDF (α)	89.06	98.87	99.62	**100.00**	92.96	79.80	98.35	99.00	99.40	86.80
+ BM25 (α)	92.64	99.25	99.62	**100.00**	95.21	82.40	97.42	98.60	99.30	88.33
+ TF–IDF (RRF)	87.92	97.92	99.43	**100.00**	91.80	76.20	97.10	98.50	99.40	84.15
+ BM25 (RRF)	89.43	98.68	99.43	**100.00**	93.28	77.60	96.30	97.80	99.05	84.92
AITeamVN/Vietnamese_Embedding_v2										
Dense	90.38	99.06	99.81	99.81	93.96	86.20	98.32	98.97	99.70	91.04
+ TF–IDF (α)	92.64	**99.62**	99.81	**100.00**	95.26	85.40	**98.42**	99.10	**99.75**	90.91
+ BM25 (α)	93.77	99.25	**100.00**	**100.00**	95.95	**87.40**	98.17	98.67	99.55	**91.67**
+ TF–IDF (RRF)	90.19	98.49	99.81	**100.00**	93.15	79.10	97.65	98.60	99.55	86.65
+ BM25 (RRF)	92.08	98.68	99.81	**100.00**	94.78	81.00	96.27	97.75	99.45	87.12

provide strong retrieval signals even in heterogeneous collections. Second, BM25 hybrids deliver the highest effectiveness. The best configuration—Vietnamese Embedding v2 combined with BM25 (α)—achieves P@1 = 89.10 and MRR@10 = 93.00, nearly a 15-point improvement over BM25 alone. Comparable gains are observed with bge-m3 + BM25 (92.41) and vietnamese-document-embedding + BM25 (90.34), confirming the robustness of lexical–semantic fusion. Third, Vietnamese-specific encoders dominate under hybridization. Both Vietnamese Embedding v2 and vietnamese-document-embedding consistently surpass commercial (text-embedding-3-large) and multilingual (bge-m3, MiniLM) models, pushing performance above 93 MRR. Overall, UIT-ViQuAD exhibits a near-saturation effect: recall at R@50 exceeds 99% for most hybrids, and top configurations converge to very high MRR scores. This suggests that cross-domain open knowledge retrieval is highly tractable when lexical precision is effectively combined with Vietnamese-specific

Table 10: Retrieval results on **Healthcare Domain** (ViNewsQA and ViMedAQA).

Domain	**ViNewsQA**					**ViMedAQA_v2**				
Method	P@1	R@10	R@20	R@50	MRR@10	P@1	R@10	R@20	R@50	MRR@10
TF–IDF	52.20	79.10	84.70	90.70	60.93	61.50	84.80	88.20	91.00	69.46
BM25	59.00	80.20	84.30	89.00	66.13	65.40	84.50	87.30	90.40	71.36
text-embedding-3-large										
Dense	49.20	76.40	81.40	88.40	58.59	80.40	95.40	97.20	98.70	85.44
+ TF–IDF (α)	62.40	84.50	89.60	94.40	69.96	80.90	**96.30**	97.80	**98.90**	86.22
+ BM25 (α)	64.70	85.90	**90.80**	**94.90**	71.83	**83.00**	96.10	**97.90**	98.70	**87.37**
+ TF–IDF (RRF)	56.20	83.30	89.40	94.70	65.25	74.70	91.70	95.20	98.70	80.63
+ BM25 (RRF)	61.00	85.30	90.20	94.40	68.74	77.70	92.10	95.00	98.70	82.16
BAAI/bge-m3										
Dense	57.60	79.00	83.40	89.50	64.72	81.20	94.10	96.70	98.60	85.60
+ TF–IDF (α)	63.90	86.40	89.50	93.80	71.41	81.30	94.40	96.60	98.80	85.95
+ BM25 (α)	65.50	87.10	90.10	93.60	72.87	82.50	94.60	96.70	98.40	86.58
+ TF–IDF (RRF)	59.90	84.60	89.50	93.90	68.10	74.70	91.40	94.80	98.20	80.76
+ BM25 (RRF)	63.70	85.20	90.00	94.10	70.63	76.80	91.30	94.70	98.20	81.59
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2										
Dense	25.50	52.20	60.00	71.30	33.54	48.00	73.50	80.70	86.50	56.19
+ TF–IDF (α)	50.70	76.80	83.10	88.40	58.94	67.00	87.10	90.60	94.00	74.05
+ BM25 (α)	49.70	77.30	82.80	88.90	58.61	69.70	87.00	90.20	94.40	75.24
+ TF–IDF (RRF)	41.70	73.40	84.70	92.30	51.25	61.40	84.20	90.40	95.20	69.16
+ BM25 (RRF)	44.40	76.10	85.60	93.30	53.95	63.90	83.80	89.90	95.00	70.27
bkai-foundation-models/vietnamese-bi-encoder										
Dense	45.50	67.50	73.20	80.70	52.39	70.10	87.20	90.40	93.60	75.68
+ TF–IDF (α)	58.80	80.70	85.40	90.50	66.41	75.70	90.90	93.10	95.80	81.10
+ BM25 (α)	59.90	80.40	85.60	91.30	66.83	77.10	90.80	92.80	95.50	82.00
+ TF–IDF (RRF)	52.70	78.80	86.40	92.50	61.72	69.90	89.30	92.60	96.10	76.88
+ BM25 (RRF)	56.90	79.60	86.80	93.40	64.12	72.80	89.30	92.60	95.80	78.34
dangvantuan/vietnamese-document-embedding										
Dense	54.90	76.80	80.50	86.20	61.77	75.50	90.60	93.00	96.30	80.86
+ TF–IDF (α)	61.70	84.30	88.30	92.70	69.22	78.30	93.10	94.70	97.10	83.73
+ BM25 (α)	64.70	85.20	88.20	93.10	71.51	79.80	92.60	94.50	97.00	84.37
+ TF–IDF (RRF)	57.90	82.30	88.20	93.50	66.03	73.00	91.10	94.10	96.70	79.39
+ BM25 (RRF)	62.70	83.80	89.80	94.00	69.44	74.70	90.40	93.50	96.40	80.03
AITeamVN/Vietnamese_Embedding_v2										
Dense	55.40	78.20	82.90	88.60	63.01	76.90	93.50	96.10	98.30	82.40
+ TF–IDF (α)	65.30	86.10	90.00	93.60	72.35	79.80	94.10	96.20	98.30	84.61
+ BM25 (α)	**68.60**	**87.30**	90.00	93.80	**74.96**	81.40	93.80	95.70	98.00	85.38
+ TF–IDF (RRF)	62.10	85.40	89.90	94.30	69.71	75.00	91.80	94.60	98.00	80.92
+ BM25 (RRF)	65.10	86.30	89.90	94.50	72.00	76.60	90.80	94.40	97.80	81.43

semantic encoders.

E.7 Rank-Based Evaluation

To enable fair and interpretable comparison, we employ a rank-based aggregation scheme. This approach highlights not only the strongest methods within each dataset, but also the most consistent performers across the entire Vietnamese retrieval benchmark.

Per-metric ranking. Let $\mathcal{K}$ denote the set of evaluation metrics (e.g., Precision@1, Recall@10, MRR@10). For each dataset $d \in \mathcal{B}$, metric $k \in \mathcal{K}$, and retrieval method $s \in \mathcal{S}$, we denote the evaluation score as

$$V_{d,k}(s) = \mathcal{M}_k^d(s), \tag{22}$$

where $\mathcal{M}_k^d(\cdot)$ is the metric-specific evaluation function. The corresponding rank is then defined

Table 11: Retrieval results on **Lifestyle & Reviews** (VlogQA and ViRe4MRC).

Dataset	**VlogQA_2**					**ViRe4MRC**				
Method	P@1	R@10	R@20	R@50	MRR@10	P@1	R@10	R@20	R@50	MRR@10
TF–IDF	13.40	34.60	47.00	67.50	19.55	3.70	17.50	23.70	35.90	7.24
BM25	18.00	39.50	45.50	60.00	23.90	6.60	20.40	26.70	36.30	10.25
text-embedding-3-large										
Dense	13.50	37.00	45.60	59.50	20.45	9.70	30.00	38.60	52.40	15.03
+ TF–IDF (α)	20.90	48.20	56.70	69.40	28.95	10.80	29.60	39.10	55.20	16.20
+ BM25 (α)	21.50	48.70	59.20	71.60	29.33	12.10	30.00	39.00	54.50	16.88
+ TF–IDF (RRF)	20.10	48.40	58.70	72.10	28.72	10.70	26.90	37.20	52.50	15.32
+ BM25 (RRF)	22.10	50.10	61.70	72.60	30.59	10.90	28.50	38.20	54.00	15.69
BAAI/bge-m3										
Dense	24.20	51.90	59.90	68.60	32.49	12.40	30.80	40.00	54.20	17.25
+ TF–IDF (α)	27.20	57.70	64.90	74.60	36.19	12.00	30.90	39.50	55.70	17.64
+ BM25 (α)	**30.50**	**59.20**	66.10	75.50	**39.20**	12.40	**31.10**	**40.80**	**56.00**	18.02
+ TF–IDF (RRF)	25.10	56.90	64.50	74.90	34.62	10.30	28.70	38.00	53.00	15.86
+ BM25 (RRF)	27.50	57.00	65.20	75.80	36.31	11.50	29.30	39.30	54.10	16.88
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2										
Dense	4.50	14.80	20.50	30.70	7.16	4.70	16.10	22.10	35.50	7.30
+ TF–IDF (α)	14.40	33.20	41.30	54.50	19.67	8.40	22.20	29.90	42.70	11.89
+ BM25 (α)	12.00	32.30	39.80	53.30	17.65	8.50	23.20	31.10	44.30	12.48
+ TF–IDF (RRF)	12.30	34.00	46.70	63.10	18.02	6.90	21.00	29.80	44.00	10.60
+ BM25 (RRF)	13.10	37.90	49.10	62.30	19.66	7.30	21.90	31.20	44.10	11.26
bkai-foundation-models/vietnamese-bi-encoder										
Dense	13.90	34.30	42.80	54.30	19.46	8.60	21.10	29.00	41.80	11.97
+ TF–IDF (α)	22.50	47.40	54.90	65.80	29.60	9.80	25.80	33.40	45.10	14.48
+ BM25 (α)	21.40	46.70	54.60	65.90	28.71	10.10	27.30	33.80	46.50	15.16
+ TF–IDF (RRF)	19.90	48.90	58.50	70.40	28.26	9.40	25.40	32.50	45.20	14.14
+ BM25 (RRF)	21.10	49.40	60.30	71.10	29.60	9.80	26.30	33.70	46.60	14.47
dangvantuan/vietnamese-document-embedding										
Dense	22.70	46.90	56.20	69.20	29.62	10.70	27.40	35.70	50.50	15.02
+ TF–IDF (α)	25.50	54.90	65.40	75.80	34.33	10.80	28.90	37.80	53.40	15.83
+ BM25 (α)	28.80	57.50	**66.90**	76.20	36.96	11.60	29.70	38.20	54.50	16.55
+ TF–IDF (RRF)	22.80	55.80	65.30	75.80	32.32	10.60	28.10	35.70	50.70	15.15
+ BM25 (RRF)	25.60	57.20	66.00	**76.60**	35.03	11.00	28.20	36.90	52.00	15.55
AITeamVN/Vietnamese_Embedding_v2										
Dense	22.30	49.00	57.50	69.00	29.92	10.60	28.60	38.70	53.50	15.48
+ TF–IDF (α)	25.80	55.70	63.40	74.10	35.04	12.20	30.60	40.00	54.90	17.40
+ BM25 (α)	29.20	57.80	65.60	74.60	37.80	**13.00**	30.80	39.30	54.80	**18.21**
+ TF–IDF (RRF)	24.10	55.20	63.20	74.80	33.67	12.20	29.40	37.10	53.20	16.98
+ BM25 (RRF)	27.10	56.40	65.20	75.50	35.80	**13.00**	29.40	38.30	53.90	17.56

in Equation (23).

$$r_{d,k}(s) = 1 + \left|\{\, s' \in \mathcal{S} : V_{d,k}(s') > V_{d,k}(s) \,\}\right|, \tag{23}$$

with ties assigned the smallest rank in their group.

Normalized rank. To compare fairly across datasets with different numbers of competing methods, ranks are normalized into $[0, 1]$, as shown in Equation (24).

$$\tilde{r}_{d,k}(s) = \frac{r_{d,k}(s) - 1}{|\mathcal{S}_d| - 1}, \tag{24}$$

where $|\mathcal{S}_d|$ denotes the number of evaluated methods on dataset d. Here, $\tilde{r}_{d,k}(s) = 0$ indicates the best performance and $\tilde{r}_{d,k}(s) = 1$ the worst.

Dataset-level aggregation. For each dataset d, the aggregate rank of method s is computed by summing its normalized ranks across metrics, as given in Equation (25).

$$R_d(s) = \sum_{k \in \mathcal{K}} \tilde{r}_{d,k}(s). \tag{25}$$

Table 12: Retrieval results on **UIT-ViQuAD** (Cross-domain Open Knowledge).

Method	P@1	R@10	R@20	R@50	MRR@10
TF–IDF	50.00	91.00	94.00	97.40	64.57
BM25	70.80	91.60	93.90	97.20	78.09
text-embedding-3-large					
Dense	71.20	92.40	96.00	97.60	78.75
+ TF–IDF (α)	79.30	97.20	98.20	99.30	86.24
+ BM25 (α)	82.90	97.00	98.70	99.70	88.46
+ TF–IDF (RRF)	74.10	96.10	98.30	99.60	82.18
+ BM25 (RRF)	77.80	95.80	98.70	99.70	84.59
BAAI/bge-m3					
Dense	80.60	96.40	98.40	99.70	86.62
+ TF–IDF (α)	83.20	98.70	99.20	**99.80**	89.48
+ BM25 (α)	88.30	99.00	**99.30**	99.70	92.41
+ TF–IDF (RRF)	77.20	97.60	99.00	**99.80**	84.95
+ BM25 (RRF)	80.80	97.30	99.00	99.70	87.15
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2					
Dense	55.90	81.30	87.70	93.30	64.15
+ TF–IDF (α)	70.20	93.90	96.60	98.70	78.74
+ BM25 (α)	75.80	95.10	97.50	98.70	82.76
+ TF–IDF (RRF)	66.30	93.20	96.50	99.20	75.45
+ BM25 (RRF)	69.00	93.60	97.30	99.20	77.59
bkai-foundation-models/vietnamese-bi-encoder					
Dense	68.00	88.40	92.10	95.00	74.62
+ TF–IDF (α)	77.40	95.00	97.20	98.20	83.93
+ BM25 (α)	81.20	96.10	97.80	98.50	86.71
+ TF–IDF (RRF)	72.20	94.70	97.80	98.90	80.16
+ BM25 (RRF)	75.70	94.90	97.80	99.10	82.50
dangvantuan/vietnamese-document-embedding					
Dense	75.70	95.60	97.40	99.10	83.00
+ TF–IDF (α)	81.40	97.50	99.00	99.50	87.45
+ BM25 (α)	85.50	98.70	**99.30**	99.60	90.34
+ TF–IDF (RRF)	76.40	97.10	98.90	99.60	83.82
+ BM25 (RRF)	79.60	97.10	99.10	99.60	86.07
AITeamVN/Vietnamese_Embedding_v2					
Dense	82.20	98.10	99.00	99.70	88.24
+ TF–IDF (α)	84.40	**99.30**	**99.30**	**99.80**	90.18
+ BM25 (α)	**89.10**	99.20	**99.30**	99.70	**93.00**
+ TF–IDF (RRF)	79.00	97.60	**99.30**	**99.80**	86.02
+ BM25 (RRF)	82.10	97.50	99.10	**99.80**	87.95

Domain and overall aggregation. Each domain corresponds to a subset $\mathcal{B}^\star \subseteq \mathcal{B}$ of datasets. The domain-level rank is then calculated in Equation (26).

$$R_{\mathcal{B}^\star}(s) = \sum_{d \in \mathcal{B}^\star} R_d(s). \tag{26}$$

Finally, the overall benchmark rank aggregates across all domains, as shown in Equation (27).

$$R(s) = \sum_{\mathcal{B}^\star} R_{\mathcal{B}^\star}(s), \qquad \bar{R}(s) = \frac{R(s)}{|\mathcal{B}|}. \tag{27}$$

Lower values of $R(s)$ or $\bar{R}(s)$ indicate stronger and more consistent performance across metrics, datasets, and domains.

Results. Table 13 summarizes the top–3 methods across domains. Several clear trends emerge.

First, BM25 hybrids dominate: across most domains, leading systems combine semantic encoders with lexical signals via α-weighted interpolation. Second, Vietnamese encoders consistently outperform global counterparts in legally grounded and cross-domain datasets, where AITeamVN's Viet-

namese Embedding v2 emerges as the most reliable overall. Third, multilingual and commercial models show domain-specific strengths: OpenAI's text-embedding-3-large leads in customer service and clinical QA, while bge-m3 performs best on conversational and lifestyle queries. Fourth, performance saturation appears in structured domains: in Legal and Cross-domain, top systems converge above 90 MRR with recall near 100%, suggesting that retrieval in standardized text is largely solved. Finally, retrieval remains difficult in informal domains: Lifestyle and Customer Service yield much lower scores, reflecting persistent challenges with short, paraphrastic, and noisy queries despite hybridization.

Together, these results reinforce the complementarity of lexical and semantic retrieval, while highlighting both the strengths of domain-adapted Vietnamese embeddings and the open challenges in more conversational, less structured settings.

F Representative Error Cases

Table 14 presents typical retrieval failures discussed in Section 4.5, illustrating error types (E1–E3) through real samples from the lowest-performing datasets. Each case includes the original Vietnamese query, its English translation, the corresponding gold passage, and a concise diagnostic comment.

Table 13: Top–3 retrieval methods per domain and overall under rank-based evaluation.

Domain	**1st Place**	**2nd Place**	**3rd Place**
Education	bge-m3 + BM25 (α)	text-embedding-3-large + BM25 (α)	bge-m3 + TF–IDF (α)
Customer Service	text-embedding-3-large + TF–IDF (α)	text-embedding-3-large + BM25 (α)	Vietnamese_Embedding_v2 + BM25 (α)
Legal	Vietnamese_Embedding_v2 + TF–IDF (α)	Vietnamese_Embedding_v2 + BM25 (α)	text-embedding-3-large + BM25 (α)
Healthcare	text-embedding-3-large + BM25 (α)	bge-m3 + BM25 (α)	Vietnamese_Embedding_v2 + BM25 (α)
Lifestyle & Reviews	bge-m3 + BM25 (α)	Vietnamese_Embedding_v2 + BM25 (α)	bge-m3 + TF–IDF (α)
Cross-domain	Vietnamese_Embedding_v2 + BM25 (α)	Vietnamese_Embedding_v2 + TF–IDF (α)	bge-m3 + BM25 (α)
Overall	Vietnamese_Embedding_v2 + BM25 (α)	bge-m3 + BM25 (α)	bge-m3 + TF–IDF (α)

Table 14: Representative failure examples from low-performing datasets. Only key fragments are shown; filler, repetitive, or truncated transcript content is omitted and denoted by "[...]".

Dataset	Example & Observation
CSConDa	**Query:** gửi tn hàng loạt giá sao ak, mình cần mỗi tính năng đó. (**English:** How much is the bulk message feature? I only need that function.) **Gold passage:** DooPage cung cấp tính năng gửi tin nhắn hàng loạt, đặc biệt hữu ích cho các doanh nghiệp muốn tiếp cận nhiều khách hàng cùng lúc trên các nền tảng mạng xã hội. [...] Điều này giúp doanh nghiệp có thể nhanh chóng làm quen và đánh giá hiệu quả của DooPage trước khi quyết định đăng ký sử dụng chính thức. (**English:** DooPage provides a bulk messaging feature [...] helping businesses evaluate its effectiveness before official registration.) → The query is highly informal and shortened with teencode ("ak" = "à không" / "ah"), whereas the gold passage adopts a formal, policy-oriented register. This stylistic and lexical divergence causes severe retrieval mismatch, as neither sparse nor dense encoders reliably bridge colloquial intent to formal documentation.
EduCoQA	**Query:** trưởng khoa là ai? (**English:** Who is the dean?) **Gold passage:** Khoa Khoa học và Kỹ thuật Máy tính\nCơ cấu nhân sự:\nBan Chủ nhiệm Khoa\nTrưởng khoa: PGS. TS. [Ẩn danh] (Email: [redacted])\nPhó Trưởng khoa: PGS. TS. [Ẩn danh]\nPhó Trưởng khoa: PGS. TS. [Ẩn danh]\nPhó Trưởng khoa: PGS. TS. [Ẩn danh] (**English:** Faculty of Computer Science and Engineering\nOrganizational Structure:\nDepartment Board\nDean: Assoc. Prof. [Name Redacted]\nVice Deans: Assoc. Prof. [Name Redacted], Assoc. Prof. [Name Redacted], Assoc. Prof. [Name Redacted].) → The question is underspecified and context-free, yielding entity ambiguity: multiple plausible targets exist, and retrievers tend to surface semantically related but pragmatically irrelevant passages.
	Query: khoa ktxd có những hb nào vậy? (**English:** What scholarships are available in Civil Engineering?) **Gold passage:** Học bổng trao đổi và học bổng toàn phần [...] với Đại học Kyoto, Hiroshima (Nhật Bản), học bổng chính phủ Đài Loan, Hàn Quốc và Nhật Bản [...] cho sinh viên ngành Kỹ thuật Cơ sở Hạ tầng. (**English:** Exchange and full scholarships [...] for students of Infrastructure Engineering.) → The query uses abbreviations ("ktxd" = "kỹ thuật xây dựng" / Civil Engineering) and informal orthography without diacritics, while the gold passage is fully formal with complete accents. Orthographic and register gaps reduce lexical overlap and weaken sparse–dense fusion.
VlogQA	**Query:** Nguyên liệu làm bánh kem có những gì? (**English:** What ingredients are used to make the cake?) **Gold passage:** [...] bột mì purple flower hay bột mì đa dụng số 11 [...] sau đó đánh đều lên [...] để trong tủ lạnh 20 phút [...] ừ ừ à à ừ ừ 3 cách thay đổi school này là 15ml [...] chị em mình nhớ lấy cái này nè để chút nữa mình đổ bấm xoay tròn... (**English:** ...purple flower flour or all-purpose flour number 11 [...] uh uh ah ah [...] remember to take this one for pouring later...) → Spoken transcripts contain heavy disfluency ("ừ ừ à à" = "uh uh ah ah"), filler words, and repetition. The lack of clear sentence boundaries harms embedding coherence and destabilizes sentence-level retrieval for semantic encoders.
ViRe4MRC	**Query:** Cảm nhận của khách hàng là gì sau khi sử dụng sản phẩm? (**English:** What are customers' impressions after using the product?) **Gold passage:** k chề vào đâu đc đáp ứng tất cả các nhu cầu [...] vô cùg mượt mà [...] ☻☻☻ (**English:** No complaints at all, meets all needs [...] super smooth performance [...] ☻.) → Reviews are informal, emotion-laden, and include teencode ("k chề" = "không chê" / "no complaints"), along with emojis that disrupt tokenization and dilute sentiment cues. These artifacts make relevance estimation intrinsically noisy.