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

**Long S. T. Nguyen**[1,2,3], **Tho T. Quan**[1,2*]

[1]URA Research Group, Ho Chi Minh City University of Technology (HCMUT), Vietnam
[2]Vietnam National University Ho Chi Minh City, Vietnam
[3]Center for AI Research (CAIR), VinUniversity, Vietnam

[*]**Correspondence:** qttho@hcmut.edu.vn

## Abstract

Large Language Models (LLMs) depend on retrieval for factual grounding in Retrieval-Augmented Generation (RAG), placing Information Retrieval (IR) at the core of modern Question Answering (QA) systems. While lexical, dense, and hybrid paradigms have been extensively benchmarked in English, their relative effectiveness for Vietnamese remains insufficiently characterized, especially under realistic multi-domain settings. Existing studies are typically confined to single domains or curated datasets, limiting cross-domain comparability and obscuring paradigm-level trade-offs. We introduce the first domain-normalized, multi-domain benchmark for Vietnamese IR under a unified and reproducible evaluation protocol, spanning six domains and ten datasets across education, legal, healthcare, customer support, lifestyle reviews, and open-domain knowledge. We evaluate lexical, neural-sparse, late-interaction, dense, and hybrid paradigms across diverse Vietnamese-specific and multilingual embedding backbones, and release two QA datasets, EduCoQA and CSConDa, constructed from authentic counseling and customer-service interactions. Beyond reporting benchmark performance, we derive systematic insights into lexical–semantic hybridization, specialization versus robustness trade-offs, and the limited predictive value of model scale for retrieval effectiveness. All datasets and evaluation scripts are publicly available at https://github.com/longstnguyen/ViRE.

## 1 Introduction

*Large Language Models* (LLMs) have become the backbone of modern conversational agents and domain-specific *Question Answering* (QA) systems (Kamalloo et al., 2023; Chang et al., 2024). Despite their impressive fluency and generalization ability, LLMs remain fundamentally constrained by their reliance on parametric knowledge (Chang et al., 2024). This limitation often manifests as hallucination, where models generate responses that are coherent but factually incorrect. *Retrieval-Augmented Generation* (RAG) mitigates this issue by grounding generation in externally retrieved documents rather than relying solely on internal model memory (Lewis et al., 2020). In practical deployments, the reliability of a RAG system is therefore determined primarily by the effectiveness of its retriever, placing *Information Retrieval* (IR) at the core of modern QA pipelines (Fan et al., 2024; Arslan et al., 2024).

A broad spectrum of IR paradigms has been developed, including lexical methods such as TF–IDF and BM25, neural-sparse approaches, late-interaction architectures, dense embedding models, and hybrid strategies that integrate lexical and semantic signals (Robertson and Zaragoza, 2009; Khattab and Zaharia, 2020; Formal et al., 2021). These paradigms have been extensively evaluated in English and other high-resource languages under large-scale and standardized benchmarks. However, their relative strengths and weaknesses in Vietnamese remain insufficiently characterized.

Vietnamese poses distinctive challenges for retrieval. Realistic user queries frequently contain abbreviations, slang, code-switching, spelling variation, and missing diacritics. Target corpora, including customer support logs, educational policy documents, and online reviews, are often noisy and stylistically heterogeneous. These properties differ substantially from the clean and standardized corpora that dominate many existing evaluations. Prior studies on Vietnamese retrieval typically focus on individual datasets or narrow domains (Ha et al., 2024; Pham Duy and Le Thanh, 2023; Nguyen et al., 2024), limiting cross-domain comparability and yielding fragmented conclusions about model robustness. As a result, it remains unclear which retrieval paradigms and encoder families provide the most reliable performance for Vietnamese in realistic multi-domain QA settings.



*Findings of the Association for Computational Linguistics: EACL 2026*, pages 2098–2128
March 24–29, 2026 

This work addresses this gap through a unified and domain-normalized benchmark for Vietnamese retrieval designed to support practical QA deployment. We investigate the following research question: *Which retrieval paradigms and encoder families offer the most robust and consistent performance for Vietnamese across diverse domains and query styles?*

To answer this question, we construct a multi-domain benchmark spanning six domains and ten datasets, covering education, legal, healthcare, customer support, lifestyle reviews, and open-domain knowledge. All datasets are standardized under a consistent and reproducible evaluation protocol to ensure fair cross-domain comparison. We systematically evaluate lexical, neural-sparse, late-interaction, dense, and hybrid retrieval paradigms across a broad range of Vietnamese-specialized and multilingual embedding backbones. This controlled experimental framework allows us to isolate paradigm-level and architectural effects while minimizing dataset-specific confounders.

Our contributions are summarized as follows.

- We establish the first domain-normalized and multi-domain benchmark for Vietnamese retrieval under a unified and reproducible evaluation protocol.
- We conduct a comprehensive cross-paradigm comparison of lexical, neural-sparse, late-interaction, dense, and hybrid retrieval methods across diverse Vietnamese-specific and multilingual embedding models.
- We introduce two realistic Vietnamese QA datasets, EduCoQA and CSConDa, constructed from authentic educational counseling and customer-service interactions that capture natural linguistic variation and noise.
- We provide systematic empirical insights into lexical and semantic hybridization, domain-dependent specialization versus robustness trade-offs, and the limited predictive value of model scale for retrieval effectiveness.

Together, these contributions deliver the first controlled and large-scale empirical study of Vietnamese retrieval across domains. Beyond benchmarking existing approaches, our findings offer practical guidance for designing robust Vietnamese RAG systems and establish a reproducible foundation for future research in low-resource and cross-lingual retrieval.

## 2 Related Work

Research on Vietnamese IR remains relatively limited and largely domain-specific. Existing work can be grouped into two primary directions: domain-bound retrieval evaluations and the construction of Vietnamese QA corpora. However, neither line provides a unified, cross-domain comparison of retrieval paradigms under realistic conditions.

**Domain-specific retrieval studies.** Several studies have explored retrieval techniques in the Vietnamese legal domain, examining lexical methods such as BM25 as well as dense embeddings within RAG-based legal QA systems (Ba et al., 2024; Ha et al., 2024; Pham Duy and Le Thanh, 2023; Khang et al., 2024). These efforts demonstrate the feasibility of retrieval-augmented approaches in specialized settings but remain confined to a single domain. More recent work investigates Vietnamese dense encoders and embedding refinement strategies (T. and T., 2024; Nguyen et al., 2024, 2025). While these studies report improvements within their respective datasets, they do not systematically compare lexical, neural-sparse, late-interaction, dense, and hybrid paradigms under a unified evaluation framework. As a result, existing findings provide domain-bound insights but do not clarify cross-domain robustness or paradigm-level trade-offs.

**Vietnamese QA datasets.** Parallel to retrieval research, several Vietnamese QA corpora have been introduced, including Wikipedia[1]-style datasets and domain-focused resources such as healthcare benchmarks (Van Nguyen et al., 2020, 2022; Tran et al., 2024). These datasets have primarily supported machine reading comprehension and fine-tuned language model evaluation rather than systematic retrieval analysis. Moreover, their texts are typically curated and well-structured, which limits their ability to capture informal queries, linguistic variation, and real-world noise commonly observed in customer support or educational counseling scenarios. Consequently, they provide limited evidence for understanding retrieval performance under realistic deployment conditions.

In summary, prior work either evaluates retrieval methods within a single domain or focuses on cu-

[1] https://en.wikipedia.org/wiki/Main_Page

rated QA benchmarks without controlled cross-paradigm comparison. To our knowledge, no study has established a domain-normalized and multi-domain benchmark that systematically evaluates lexical, neural-sparse, late-interaction, dense, and hybrid retrieval methods for Vietnamese under a unified and reproducible protocol. Our work addresses this gap by providing a controlled cross-domain evaluation framework and introducing two realistic datasets that reflect authentic Vietnamese usage scenarios.

## 3 Methodology

### 3.1 RAG Formulation

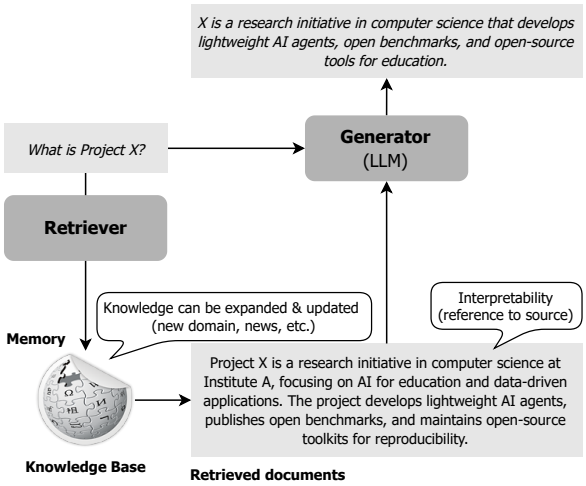


Figure 1: Overview of RAG. The retriever selects relevant documents from a knowledge base, which are then provided to the generator (LLM) as additional context for answer generation.

RAG decomposes question answering into two interacting components: a retriever and a generator, as illustrated in Figure 1. Given a query $q$ and a document collection $\mathcal{D} = \{d_i\}_{i=1}^{N}$, the retriever assigns a relevance score $s(q, d)$ to each document and returns the top-$k$ candidates:

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

Let $L_\theta$ denote the generator, typically instantiated as a LLM. Conditioned on the query and the retrieved evidence $\mathcal{S}_k$, the generator produces an output sequence by maximizing the conditional likelihood:

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

When retrieval is omitted ($k = 0$), the formulation reduces to a standard conditional language model:

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

Equation (1) and Equation (2) make explicit that the generator can only reason over the retrieved evidence $\mathcal{S}_k$; consequently, missing or irrelevant documents directly constrain downstream reasoning and answer quality.

**Retrieval as the Structural Bottleneck.** In RAG systems, retrieval defines the upper bound of achievable performance.

This dependence is particularly consequential for Vietnamese, where queries are often informal, abbreviated, or code-switched, and documents may exhibit substantial stylistic variation or noise. Under such conditions, retrieval errors are amplified and propagate directly to the final generation output, motivating careful evaluation of retrieval strategies.

### 3.2 Retrieval Methods

Equation (1) defines retrieval as ranking documents via a relevance function $s(q, d)$, and Equation (8) specifies the objective of selecting the scoring function that maximizes evaluation performance. Accordingly, our study benchmarks alternative instantiations of $s(q, d)$ for Vietnamese retrieval.

We compare a broad spectrum of retrieval paradigms, including lexical methods (Salton and Buckley, 1988; Robertson and Zaragoza, 2009), neural-sparse retrievers (Formal et al., 2021), late-interaction models (Khattab and Zaharia, 2020), dense retrievers with Vietnamese-specific and multilingual embedding backbones, and hybrid schemes that fuse sparse and dense signals via $\alpha$-weighted interpolation and *Reciprocal Rank Fusion* (RRF) (Cormack et al., 2009). The dense backbones span conventional Transformer-based encoders and recent LLM-based embedding models, enabling comprehensive analysis across architectures, fusion strategies, and Vietnamese domains.

### 3.3 Benchmarking Setup

We cast retrieval benchmarking as an optimization problem over scoring functions. Let $s(q, d)$ denote a retrieval scoring function that assigns a relevance score to a query–document pair $(q, d)$.

For each dataset, we define a benchmark as a set

of retrieval instances:

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

where $q_j$ is a query, $RD_j \subseteq \mathcal{D}$ denotes the set of ground-truth relevant documents for $q_j$, and $\mathcal{D}$ is the document corpus of the dataset. In our benchmark suite, each *domain* is represented by one or two datasets that capture its characteristic query styles and document distributions. Results are reported both at the domain level and across all domains.

Given a scoring function $s$, retrieval for query $q_j$ selects the top-$k$ documents as:

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

which instantiates the ranking operator defined in Equation (1).

Let $\mathcal{M}(\cdot)$ denote a retrieval evaluation metric, such as Precision@$k$ or Recall@$k$ (Manning et al., 2008). The performance of a scoring function $s$ on benchmark $\mathcal{B}$ is defined as the average metric value across all queries:

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

We evaluate retrieval performance at two levels. For *domain-wise* analysis, $\mathrm{Score}_{\mathcal{B}}(s)$ is computed separately for each domain-specific benchmark. For *overall* evaluation, domain scores are aggregated as:

$$\mathrm{Score}_{\mathrm{all}}(s) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} \mathrm{Score}_{\mathcal{B}^{(g)}}(s), \tag{7}$$

where $\mathcal{G}$ denotes the set of domains and $\mathcal{B}^{(g)}$ is the benchmark associated with domain $g$.

The benchmarking objective is to identify the best-performing retrieval configuration within a candidate set $\mathcal{S}$:

$$s^* = \arg\max_{s \in \mathcal{S}} \mathrm{Score}_{\mathrm{all}}(s), \tag{8}$$

where $\mathcal{S}$ includes lexical methods, sparse neural retrievers, late-interaction models, dense embedding-based retrievers, and hybrid fusion strategies.

This formulation yields a model-agnostic and domain-aware evaluation protocol, ensuring that observed performance differences can be attributed to the retrieval paradigm and embedding backbone rather than dataset- or protocol-specific confounding factors.

# 4 Experiments

We benchmark Vietnamese information retrieval under a unified multi-domain setup, comparing lexical, neural-sparse, late-interaction, dense, and hybrid paradigms across six domains, with an emphasis on cross-domain robustness and practical guidance for QA deployment.

## 4.1 Evaluation Metrics

Retrieval effectiveness is assessed using four standard metrics: Precision@$k$, Recall@$k$, MRR@$k$, and nDCG@$k$ (Yu et al., 2025). All metrics are defined at the query level and subsequently aggregated across queries as specified in Equation (6). Here, $\mathbf{1}[\cdot]$ denotes the indicator function, which equals 1 if the condition holds and 0 otherwise.

**Precision@$k$.** *Precision@$k$* (P@$k$) measures the proportion of retrieved documents that are relevant:

$$\mathrm{Precision@}k(q_j) = \frac{\left|\mathcal{S}_k^j(s) \cap RD_j\right|}{k}. \tag{9}$$

**Recall@$k$.** *Recall@$k$* (R@$k$) measures the fraction of relevant documents retrieved within the top-$k$ results:

$$\mathrm{Recall@}k(q_j) = \frac{\left|\mathcal{S}_k^j(s) \cap RD_j\right|}{|RD_j|}. \tag{10}$$

**MRR@$k$.** *Mean Reciprocal Rank* (MRR@$k$) evaluates ranking fidelity by rewarding systems that return a relevant document at higher ranks. Let $\mathrm{rank}_j$ denote the rank position of the first relevant document in $\mathcal{S}_k^j(s)$ for query $q_j$.

$$\mathrm{MRR@}k(q_j) = \frac{\mathbf{1}[\mathrm{rank}_j \le k]}{\mathrm{rank}_j}. \tag{11}$$

**nDCG@$k$.** *Normalized Discounted Cumulative Gain* (nDCG@$k$) evaluates overall ranking quality by assigning higher importance to relevant documents appearing at higher ranks. Let $d_i$ denote the document ranked at position $i$ in $\mathcal{S}_k^j(s)$. For query $q_j$, the Discounted Cumulative Gain is defined as:

$$\mathrm{DCG@}k(q_j) = \sum_{i=1}^{k} \frac{\mathbf{1}[d_i \in RD_j]}{\log_2(i+1)}, \tag{12}$$

where $\mathrm{IDCG@}k(q_j)$ denotes the DCG@$k$ of the ideal ranking in which all relevant documents for $q_j$ are ranked first. The normalized score is:

$$\mathrm{nDCG@}k(q_j) = \frac{\mathrm{DCG@}k(q_j)}{\mathrm{IDCG@}k(q_j)}. \tag{13}$$

### 4.2 Datasets

We construct a multi-domain Vietnamese retrieval benchmark in which naturally occurring queries are paired with domain-specific documents, promoting linguistic diversity and realistic retrieval difficulty. Details of our newly proposed datasets, *EduCoQA* and *CSConDa*, are provided in Appendix D, while comprehensive statistics and analyses for all benchmarks appear in Appendix C.

**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)[2], and (ii) the *Zalo Legal Text Retrieval Challenge*[3]. 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.** To provide broad coverage, we use *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 introducing open-domain queries.

**Sampling and Normalization.** To enable fair cross-domain comparison, each dataset is standardized to 1,000 query–document pairs. We remove duplicate contexts, retain one query per unique context, and remap gold relevance labels accordingly. This fixed-size sampling ensures consistency across domains and shifts the evaluation focus toward linguistic variability rather than corpus size effects, following the multi-domain spirit of benchmarks such as BeIR (Thakur et al., 2021). We additionally report a full-corpus validation study in Appendix F.1 to confirm that our conclusions remain unchanged.

### 4.3 Baselines

Building on the retrieval formulation introduced in Section 3.2, we instantiate a comprehensive set of retrieval systems for empirical comparison. Unless otherwise specified, all models are evaluated in their original pretrained form under a unified experimental protocol. Detailed implementation settings and infrastructure are described in Appendix A.

**Lexical.** We implement TF–IDF (Salton and Buckley, 1988) and BM25 (Robertson and Zaragoza, 2009) using standard inverted-index configurations, serving as non-neural reference baselines grounded in term statistics.

**Neural-sparse.** We adopt SPLADE (Formal et al., 2021), which generates vocabulary-aligned sparse representations from contextualized token embeddings, enabling term expansion while remaining compatible with inverted indexing.

**Late interaction.** We evaluate ColBERT (Khattab and Zaharia, 2020), which independently encodes queries and documents and computes relevance via token-level similarity aggregation.

**Dense.** We evaluate a diverse set of publicly available embedding models spanning Vietnamese-specialized encoders, multilingual bi-encoders, long-context architectures, and LLM-derived models. Dense retrieval is performed using cosine similarity over normalized embeddings. Model specifications are summarized in Table 1, with additional architectural details provided in Appendix B.

**Hybrid.** Hybrid configurations combine lexical and dense signals using (i) linear score interpolation with a tunable $\alpha$ parameter and (ii) RRF.

[2] https://alqac.github.io

[3] https://challenge.zalo.ai/portal/legal-text-retrieval

Table 1: Neural retrieval models evaluated in this study, including dense, late-interaction (ColBERT), and learned-sparse (SPLADE) architectures. We report backbone parameter scale and embedding dimensionality.

| Model ID | #Params | Embedding Dim. |
|---|---|---|
| OpenAI's text-embedding-3-large | – | 3,072 |
| AITeamVN/Vietnamese_Embedding_v2 | 567,754,752 | 1,024 |
| bkai-foundation-models/vietnamese-bi-encoder | 134,998,272 | 768 |
| dangvantuan/vietnamese-document-embedding | 305,368,320 | 768 |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | 117,653,760 | 384 |
| intfloat/multilingual-e5-large | 559,890,432 | 1,024 |
| jinaai/jina-embeddings-v3 | 567,754,752 | 1,024 |
| BAAI/bge-m3 | 567,754,752 | 1,024 |
| Snowflake/snowflake-arctic-embed-l-v2.0 | 567,754,752 | 1,024 |
| Alibaba-NLP/gte-multilingual-base | 305,368,320 | 768 |
| BAAI/bge-multilingual-gemma2 | 9,241,713,152 | 3,584 |
| google/EmbeddingGemma-300m | 302,863,104 | 768 |
| Alibaba-NLP/gte-Qwen2-1.5B-instruct | 1,543,268,864 | 1,536 |
| Qwen/Qwen3-Embedding-0.6B | 595,776,512 | 1,024 |
| colbert-ir/colbertv2.0 | 109,482,240 | 768 |
| naver/splade-v3 | 109,482,240 | 768 |

Table 2: Retrieval results on Education (EduCoQA, ViRHE4QA) and Customer Support (CSConDa). All values are reported as percentages (%). Best results are shown in **bold**, and second-best results are underlined.

| Domain | Education | | | | | | | | | | Customer Support | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | EduCoQA | | | | | ViRHE4QA | | | | | CSConDa | | | | |
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| TF–IDF | 14.68 | 42.47 | 22.23 | 26.99 | 53.82 | 55.60 | 92.00 | 67.70 | 73.60 | 95.90 | 15.70 | 38.50 | 22.49 | 26.30 | 47.20 |
| BM25 | 14.68 | 43.44 | 23.09 | 27.93 | 53.42 | 65.80 | 93.50 | 76.05 | 80.34 | 96.90 | 17.40 | 36.80 | 22.99 | 26.27 | 45.90 |
| ColBERT | 14.48 | 39.33 | 21.54 | 25.75 | 48.73 | 42.90 | 73.40 | 52.49 | 57.50 | 81.80 | 11.20 | 28.30 | 15.79 | 18.72 | 35.00 |
| SPLADE | 11.35 | 30.53 | 16.62 | 19.91 | 39.92 | 35.40 | 72.90 | 46.57 | 52.83 | 82.20 | 9.10 | 22.20 | 12.65 | 14.89 | 27.70 |
| text-embedding-3-large | | | | | | | | | | | | | | | |
| Dense | 20.16 | 51.86 | 30.30 | 35.47 | 63.01 | 52.60 | 88.80 | 64.94 | 70.74 | 93.60 | 33.70 | 56.80 | 41.06 | 44.84 | 63.80 |
| + BM25 ($\alpha$) | 22.11 | 57.34 | 32.54 | 38.43 | 65.95 | 66.70 | 95.40 | 76.74 | 81.30 | 97.90 | **36.40** | 60.20 | **43.60** | 47.55 | 66.20 |
| AITeamVN/Vietnamese_Embedding_v2 | | | | | | | | | | | | | | | |
| Dense | 19.96 | 50.29 | 29.11 | 34.17 | 59.88 | 61.40 | 92.20 | 72.04 | 76.96 | 95.60 | 31.40 | 54.00 | 38.40 | 42.14 | 61.40 |
| + BM25 ($\alpha$) | 19.57 | 56.36 | 29.95 | 36.20 | 65.17 | **72.50** | **96.90** | **81.42** | **85.23** | **98.80** | 33.70 | 57.90 | 41.22 | 45.20 | 64.30 |
| + BM25 (RRF) | 21.14 | 53.42 | 29.94 | 35.47 | 64.58 | 68.60 | 96.20 | 78.72 | 83.03 | 98.30 | 28.80 | 54.70 | 36.70 | 40.99 | 62.40 |
| bkai-foundation-models/vietnamese-bi-encoder | | | | | | | | | | | | | | | |
| Dense | 18.79 | 48.53 | 27.01 | 32.07 | 58.51 | 46.80 | 77.80 | 56.58 | 61.67 | 85.50 | 15.70 | 34.90 | 21.09 | 24.34 | 41.70 |
| dangvantuan/vietnamese-document-embedding | | | | | | | | | | | | | | | |
| Dense | 20.55 | 53.42 | 30.76 | 36.19 | 63.21 | 50.80 | 86.70 | 63.01 | 68.75 | 92.20 | 28.40 | 53.00 | 36.12 | 40.18 | 59.90 |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | | | | | | | | | | | | | | | |
| Dense | 14.48 | 40.70 | 20.90 | 25.49 | 52.05 | 30.80 | 63.00 | 40.43 | 45.81 | 71.90 | 11.80 | 30.00 | 16.71 | 19.82 | 39.10 |
| intfloat/multilingual-e5-large | | | | | | | | | | | | | | | |
| Dense | 23.87 | 53.82 | 32.70 | 37.72 | 65.56 | 58.50 | 91.90 | 69.85 | 75.21 | 95.60 | 27.20 | 48.10 | 33.78 | 37.21 | 55.90 |
| + TF–IDF ($\alpha$) | 23.29 | **61.06** | 33.62 | 40.08 | 68.88 | 65.40 | 94.50 | 75.96 | 80.52 | 97.60 | 31.90 | 53.60 | 38.76 | 42.32 | 61.70 |
| + BM25 ($\alpha$) | **25.24** | 60.47 | **35.24** | **41.19** | **70.06** | 69.60 | 95.50 | 79.19 | 83.21 | 97.70 | 32.60 | 54.90 | 39.46 | 43.16 | 60.70 |
| jinaai/jina-embeddings-v3 | | | | | | | | | | | | | | | |
| Dense | 21.72 | 52.05 | 31.01 | 36.02 | 62.43 | 49.50 | 86.10 | 61.08 | 67.09 | 91.30 | 32.10 | 57.40 | 40.03 | 44.19 | 64.30 |
| + TF–IDF ($\alpha$) | 22.11 | 58.32 | 32.34 | 38.49 | 68.30 | 59.60 | 92.30 | 70.88 | 76.10 | 96.50 | 34.90 | **61.20** | 42.70 | 47.12 | 66.90 |
| + BM25 ($\alpha$) | 23.48 | 59.30 | 34.13 | 40.12 | 67.91 | 64.10 | 93.70 | 74.08 | 78.83 | 97.00 | 35.40 | **61.20** | 43.42 | **47.68** | **67.60** |
| BAAI/bge-m3 | | | | | | | | | | | | | | | |
| Dense | 24.66 | 55.77 | 34.22 | 39.37 | 64.77 | 59.20 | 91.10 | 70.40 | 75.45 | 95.10 | 30.80 | 53.90 | 37.98 | 41.80 | 61.00 |
| + BM25 ($\alpha$) | 22.90 | 58.71 | 33.68 | 39.63 | 67.51 | 71.10 | 95.90 | 79.94 | 83.86 | 98.20 | 33.90 | 56.90 | 40.97 | 44.78 | 63.90 |
| Snowflake/snowflake-arctic-embed-l-v2.0 | | | | | | | | | | | | | | | |
| Dense | 20.74 | 54.79 | 31.05 | 36.73 | 64.58 | 60.00 | 91.50 | 70.41 | 75.50 | 95.30 | 32.80 | 56.70 | 40.69 | 44.56 | 63.20 |
| + BM25 ($\alpha$) | 23.48 | 58.51 | 33.33 | 39.28 | 68.49 | 68.20 | 95.70 | 78.02 | 82.35 | 97.70 | 35.60 | 59.80 | 43.56 | 47.46 | 66.30 |
| Alibaba-NLP/gte-multilingual-base | | | | | | | | | | | | | | | |
| Dense | 18.00 | 52.45 | 28.72 | 34.42 | 62.82 | 54.40 | 87.80 | 65.52 | 70.91 | 93.10 | 28.10 | 51.60 | 35.23 | 39.13 | 57.70 |
| BAAI/bge-multilingual-gemma2 | | | | | | | | | | | | | | | |
| Dense | 12.13 | 40.90 | 20.03 | 24.95 | 51.08 | 50.80 | 82.70 | 61.73 | 66.84 | 89.10 | 14.30 | 30.50 | 18.65 | 21.43 | 37.00 |
| google/EmbeddingGemma-300m | | | | | | | | | | | | | | | |
| Dense | 21.14 | 53.82 | 30.70 | 36.19 | 62.23 | 55.20 | 88.50 | 66.37 | 71.73 | 92.80 | 29.90 | 54.70 | 37.34 | 41.49 | 60.80 |
| Alibaba-NLP/gte-Qwen2-1.5B-instruct | | | | | | | | | | | | | | | |
| Dense | 6.26 | 21.92 | 9.98 | 12.75 | 29.55 | 31.90 | 68.60 | 42.71 | 48.87 | 77.30 | 6.70 | 16.50 | 9.33 | 11.01 | 21.20 |
| Qwen/Qwen3-Embedding-0.6B | | | | | | | | | | | | | | | |
| Dense | 21.53 | 54.99 | 31.50 | 37.10 | 65.95 | 55.40 | 87.80 | 66.28 | 71.48 | 93.20 | 24.80 | 49.40 | 32.27 | 36.34 | 56.60 |
| + TF–IDF ($\alpha$) | 21.72 | 59.69 | 32.47 | 38.92 | 69.67 | 60.80 | 93.00 | 72.12 | 77.22 | 96.90 | 29.10 | 56.40 | 37.31 | 41.85 | 63.10 |

### 4.4 Results and Analysis

Tables 2, 3, and 4 report multi-domain retrieval performance across all ten datasets. We include full results for lexical (TF–IDF, BM25), neural-sparse (SPLADE), late-interaction (ColBERT), and dense retrievers. For hybrid methods, only the strongest variant per backbone is shown in the main tables for readability, with complete results deferred to Ap-

Table 3: Retrieval results on Legal (ALQAC, ZaloLegalQA) and Healthcare (ViNewsQA, ViMedAQA).

| **Domain** | **Legal** | | | | | | | | | | **Healthcare** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | ALQAC | | | | | ZaloLegalQA | | | | | ViNewsQA | | | | | ViMedAQA | | | | |
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| TF–IDF | 82.83 | 96.23 | 88.34 | 90.31 | 98.11 | 64.70 | 92.47 | 75.07 | 79.28 | 96.45 | 52.20 | 79.10 | 60.93 | 65.31 | 84.70 | 61.50 | 84.80 | 69.46 | 73.18 | 88.20 |
| BM25 | 89.25 | 97.92 | 92.20 | 93.59 | 99.25 | 71.40 | 92.18 | 79.39 | 82.49 | 94.42 | 59.00 | 80.20 | 66.13 | 69.53 | 84.30 | 65.40 | 84.50 | 71.36 | 74.51 | 87.30 |
| ColBERT | 69.43 | 89.62 | 76.42 | 79.63 | 93.40 | 54.40 | 75.98 | 61.78 | 65.07 | 80.73 | 29.70 | 51.00 | 35.77 | 39.37 | 57.50 | 49.60 | 71.40 | 56.42 | 60.01 | 75.50 |
| SPLADE | 68.68 | 91.32 | 76.33 | 79.97 | 95.66 | 40.30 | 65.53 | 48.42 | 52.35 | 72.58 | 26.70 | 50.20 | 33.44 | 37.40 | 58.00 | 47.40 | 68.80 | 54.08 | 57.60 | 73.60 |
| text-embedding-3-large | | | | | | | | | | | | | | | | | | | | |
| Dense | 84.53 | 98.68 | 90.13 | 92.26 | 99.81 | 80.50 | 96.83 | 87.10 | 89.39 | 98.53 | 49.20 | 76.40 | 58.59 | 62.90 | 81.40 | 80.40 | 95.40 | 85.44 | 87.84 | 97.20 |
| + TF–IDF ($\alpha$) | 88.87 | 99.25 | 93.22 | 94.74 | 99.81 | 82.20 | 98.07 | 88.52 | 90.80 | 99.27 | 62.40 | 84.50 | 69.96 | 73.49 | 89.60 | 80.90 | 96.30 | 86.22 | 88.66 | 97.80 |
| + BM25 ($\alpha$) | 93.02 | 99.43 | 95.79 | 96.72 | **100.00** | 84.00 | 97.93 | 89.83 | 91.77 | 98.92 | 64.70 | 85.90 | 71.83 | 75.23 | 90.80 | 83.00 | 96.10 | 87.37 | 89.48 | 97.90 |
| AITeamVN/Vietnamese_Embedding_v2 | | | | | | | | | | | | | | | | | | | | |
| Dense | 90.38 | 99.06 | 93.96 | 95.24 | 99.81 | 86.20 | 98.32 | 91.04 | 92.74 | 98.97 | 55.40 | 78.20 | 63.01 | 66.69 | 82.90 | 76.90 | 93.50 | 82.40 | 85.06 | 96.10 |
| + BM25 ($\alpha$) | 93.77 | 99.25 | 95.95 | 96.78 | **100.00** | **87.40** | 98.17 | **91.67** | **93.18** | 98.67 | **68.60** | 87.30 | **74.96** | **77.95** | 90.00 | 81.40 | 93.80 | 85.38 | 87.40 | 95.70 |
| bkai-foundation-models/vietnamese-bi-encoder | | | | | | | | | | | | | | | | | | | | |
| Dense | 80.75 | 95.66 | 86.21 | 88.52 | 98.11 | 71.00 | 92.72 | 79.20 | 82.46 | 94.72 | 45.50 | 67.50 | 52.39 | 56.01 | 73.20 | 70.10 | 87.20 | 75.68 | 78.46 | 90.40 |
| dangvantuan/vietnamese-document-embedding | | | | | | | | | | | | | | | | | | | | |
| Dense | 85.85 | 98.49 | 90.51 | 92.48 | 99.25 | 77.70 | 96.27 | 85.08 | 87.81 | 97.95 | 54.90 | 76.80 | 61.77 | 65.38 | 80.50 | 75.50 | 90.60 | 80.86 | 83.24 | 93.00 |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | | | | | | | | | | | | | | | | | | | | |
| Dense | 65.85 | 92.45 | 74.37 | 78.73 | 95.09 | 51.10 | 80.23 | 60.75 | 65.30 | 85.87 | 25.50 | 52.20 | 33.54 | 37.98 | 60.00 | 48.00 | 73.50 | 56.19 | 60.34 | 80.70 |
| intfloat/multilingual-e5-large | | | | | | | | | | | | | | | | | | | | |
| Dense | 89.06 | 99.25 | 92.91 | 94.48 | 99.81 | 83.80 | 97.98 | 89.69 | 91.64 | 98.88 | 57.40 | 79.70 | 64.91 | 68.48 | 84.40 | 83.30 | 96.40 | 87.94 | 90.01 | 97.80 |
| + BM25 ($\alpha$) | 92.64 | 99.43 | 95.47 | 96.47 | **100.00** | 84.80 | 98.08 | 90.15 | 92.05 | 98.53 | 66.10 | **88.50** | 73.59 | 77.19 | **91.60** | **84.70** | **96.50** | **88.82** | **90.69** | 97.70 |
| + BM25 (RRF) | 91.32 | 99.25 | 94.73 | 95.87 | 99.81 | 78.20 | 95.33 | 84.86 | 87.41 | 97.57 | 65.00 | 87.00 | 72.10 | 75.68 | 91.40 | 81.20 | 94.60 | 85.92 | 88.03 | 96.50 |
| jinaai/jina-embeddings-v3 | | | | | | | | | | | | | | | | | | | | |
| Dense | 84.34 | 98.49 | 89.64 | 91.83 | 99.62 | 83.10 | 98.85 | 89.39 | 91.67 | **99.40** | 55.00 | 76.40 | 61.95 | 65.44 | 81.70 | 79.90 | 95.40 | 85.29 | 87.75 | 97.30 |
| + TF–IDF ($\alpha$) | 87.92 | 99.06 | 92.65 | 94.27 | **100.00** | 81.40 | **98.90** | 88.36 | 90.94 | 99.25 | 62.40 | 83.50 | 69.46 | 72.86 | 88.20 | 82.50 | 95.80 | 87.39 | 89.45 | 97.80 |
| + BM25 ($\alpha$) | 90.38 | 99.62 | 94.12 | 95.49 | 99.81 | 85.00 | 98.55 | 90.43 | 92.40 | 99.15 | 64.80 | 84.90 | 71.60 | 74.83 | 90.00 | **84.70** | 96.00 | 88.66 | 90.45 | 97.70 |
| + TF–IDF (RRF) | 85.28 | 99.06 | 90.99 | 93.01 | **100.00** | 76.40 | 97.75 | 84.77 | 87.95 | 98.90 | 58.30 | 83.10 | 66.57 | 70.57 | 88.70 | 77.10 | 94.10 | 83.83 | 86.38 | 95.90 |
| BAAI/bge-m3 | | | | | | | | | | | | | | | | | | | | |
| Dense | 90.38 | 99.43 | 94.14 | 95.47 | **100.00** | 82.30 | 97.83 | 88.52 | 90.75 | 98.77 | 57.60 | 79.00 | 64.72 | 68.17 | 83.40 | 81.20 | 94.10 | 85.60 | 87.67 | 96.70 |
| + BM25 ($\alpha$) | **94.72** | 99.62 | **96.66** | **97.40** | 99.81 | 82.10 | 98.02 | 88.62 | 90.93 | 98.62 | 65.50 | 87.10 | 72.87 | 76.33 | 90.10 | 82.50 | 94.60 | 86.58 | 88.52 | 96.70 |
| Snowflake/snowflake-arctic-embed-l-v2.0 | | | | | | | | | | | | | | | | | | | | |
| Dense | 88.30 | 99.43 | 92.68 | 94.36 | 99.81 | 83.20 | 98.03 | 89.30 | 91.42 | 98.72 | 52.70 | 75.00 | 60.12 | 63.71 | 79.70 | 79.40 | 94.30 | 84.56 | 86.93 | 96.00 |
| Alibaba-NLP/gte-multilingual-base | | | | | | | | | | | | | | | | | | | | |
| Dense | 88.30 | 98.30 | 92.04 | 93.59 | 99.62 | 79.00 | 96.22 | 85.56 | 88.13 | 98.32 | 53.00 | 76.60 | 60.79 | 64.60 | 80.40 | 73.90 | 89.90 | 79.53 | 82.05 | 92.80 |
| + TF–IDF (RRF) | 88.11 | 98.68 | 92.06 | 93.70 | **100.00** | 74.20 | 97.05 | 82.97 | 86.40 | 98.25 | 58.30 | 82.40 | 65.97 | 69.92 | 88.20 | 74.60 | 91.60 | 80.99 | 83.61 | 94.50 |
| BAAI/bge-multilingual-gemma2 | | | | | | | | | | | | | | | | | | | | |
| Dense | 85.28 | 99.06 | 90.37 | 92.51 | 99.43 | 75.60 | 95.68 | 83.21 | 86.18 | 97.15 | 44.00 | 70.40 | 52.36 | 56.68 | 75.00 | 58.40 | 80.20 | 65.39 | 68.94 | 86.50 |
| google/EmbeddingGemma-300m | | | | | | | | | | | | | | | | | | | | |
| Dense | 87.74 | 99.06 | 92.07 | 93.80 | 99.43 | 82.50 | 96.67 | 88.11 | 90.12 | 97.87 | 54.30 | 77.90 | 62.21 | 65.99 | 82.70 | 77.80 | 95.30 | 83.62 | 86.45 | 97.30 |
| + TF–IDF ($\alpha$) | 89.25 | 99.43 | 93.20 | 94.74 | 99.81 | 80.60 | 97.87 | 87.42 | 89.94 | 98.57 | 62.40 | 83.80 | 69.50 | 72.95 | 88.10 | 81.20 | 96.10 | 86.50 | 88.85 | **98.20** |
| + BM25 ($\alpha$) | 91.32 | **100.00** | 94.61 | 95.93 | **100.00** | 83.00 | 97.37 | 88.65 | 90.71 | 98.22 | 65.50 | 85.80 | 72.45 | 75.69 | 89.50 | 84.10 | 96.10 | 88.27 | 90.18 | **98.20** |
| Alibaba-NLP/gte-Qwen2-1.5B-instruct | | | | | | | | | | | | | | | | | | | | |
| Dense | 65.47 | 89.62 | 73.38 | 77.29 | 94.34 | 84.10 | 98.23 | 89.76 | 91.77 | 99.05 | 54.00 | 75.30 | 61.21 | 64.63 | 80.10 | 59.90 | 81.20 | 66.93 | 70.37 | 84.70 |
| + BM25 ($\alpha$) | 87.92 | 97.55 | 91.67 | 93.13 | 98.11 | 86.00 | 98.47 | 90.99 | 92.76 | 98.95 | 65.60 | 84.30 | 71.77 | 74.80 | 88.80 | 71.50 | 88.50 | 76.95 | 79.73 | 90.90 |
| Qwen/Qwen3-Embedding-0.6B | | | | | | | | | | | | | | | | | | | | |
| Dense | 85.66 | 99.06 | 90.44 | 92.54 | 99.43 | 80.80 | 97.68 | 87.48 | 89.88 | 98.40 | 55.00 | 76.30 | 61.97 | 65.43 | 81.40 | 77.50 | 93.00 | 82.75 | 85.23 | 95.40 |
| + BM25 ($\alpha$) | 91.51 | 99.81 | 94.60 | 95.88 | 99.81 | 83.20 | 97.63 | 88.90 | 90.96 | 98.45 | 64.60 | 85.20 | 71.44 | 74.77 | 89.80 | 82.80 | 94.80 | 87.12 | 89.00 | 96.60 |

pendix E. Evaluation uses P@1, R@10, MRR@10, nDCG@10, and R@20, reflecting practical QA-oriented retrieval settings. We summarize the principal empirical findings below.

**I1.** Lexical–semantic integration emerges as the most reliable retrieval paradigm, with linear BM25-based $\alpha$-fusion consistently outperforming standalone lexical or dense retrieval across all domains.

**I2.** Hybridization yields larger relative gains for weaker dense encoders, substantially compressing performance disparities among dense backbones, particularly on more challenging informal domains.

**I3.** Hybrid effectiveness critically depends on calibrated lexical weighting, as BM25-based linear interpolation consistently surpasses TF–IDF and rank-based fusion strategies; optimal $\alpha$ values typically lie between 0.6 and 0.8, indicating a dominant yet complementary contribution from dense similarity.

**I4.** Across encoder families, multilingual models exhibit the strongest average cross-domain robustness, with multilingual-e5-large and bge-m3 achieving the best overall performance under unified evaluation.

**I5.** Vietnamese-specialized encoders consistently excel on formal institutional corpora but generalize less effectively to highly conversational and lifestyle-oriented domains, with Vietnamese_Embedding_v2 representing the strongest performer within this group, highlighting a specialization–robustness trade-off.

**I6.** LLM-derived embedding models display pronounced cross-domain variability, as some large instruction-tuned models behave inconsistently across structured and informal datasets, whereas smaller variants often exhibit more stable behavior.

**I7.** Model scale alone does not explain retrieval quality, as larger parameter counts do not systematically yield stronger multi-domain performance.

Table 4: 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 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| TF–IDF | 13.40 | 34.60 | 19.55 | 23.10 | 47.00 | 3.70 | 17.50 | 7.24 | 9.63 | 23.70 | 50.00 | 91.00 | 64.57 | 71.05 | 94.00 |
| BM25 | 18.00 | 39.50 | 23.90 | 27.57 | 45.50 | 6.60 | 20.40 | 10.25 | 12.62 | 26.70 | 70.80 | 91.60 | 78.09 | 81.38 | 93.90 |
| ColBERT | 5.30 | 15.40 | 8.01 | 9.74 | 19.40 | 4.80 | 15.90 | 7.66 | 9.59 | 21.40 | 50.90 | 75.10 | 58.26 | 62.29 | 81.30 |
| SPLADE | 2.90 | 10.00 | 4.66 | 5.90 | 15.90 | 5.20 | 13.00 | 7.40 | 8.72 | 19.20 | 44.30 | 73.10 | 53.57 | 58.27 | 79.00 |
| text-embedding-3-large | | | | | | | | | | | | | | | |
| Dense | 13.50 | 37.00 | 20.45 | 24.37 | 45.60 | 9.70 | 30.00 | 15.03 | 18.54 | 38.60 | 71.20 | 92.40 | 78.75 | 82.09 | 96.00 |
| AITeamVN/Vietnamese_Embedding_v2 | | | | | | | | | | | | | | | |
| Dense | 22.20 | 49.00 | 29.86 | 34.39 | 57.50 | 10.60 | 28.60 | 15.48 | 18.56 | 38.70 | 82.20 | 98.10 | 88.24 | 90.67 | 99.00 |
| + TF–IDF ($\alpha$) | 25.60 | 55.80 | 34.98 | 39.97 | 63.40 | 12.20 | 30.60 | 17.40 | 20.51 | 40.00 | 84.40 | **99.30** | 90.18 | 92.45 | 99.30 |
| + BM25 ($\alpha$) | 29.20 | 57.40 | 37.73 | 42.42 | 65.90 | **13.00** | 30.80 | **18.21** | **21.19** | 39.30 | 89.10 | 99.20 | 93.00 | 94.53 | 99.30 |
| + BM25 (RRF) | 27.30 | 56.30 | 35.88 | 40.72 | 64.90 | **13.00** | 29.40 | 17.56 | 20.36 | 38.30 | 82.10 | 97.50 | 87.95 | 90.32 | 99.10 |
| bkai-foundation-models/vietnamese-bi-encoder | | | | | | | | | | | | | | | |
| Dense | 13.90 | 34.30 | 19.46 | 22.95 | 42.80 | 8.60 | 21.10 | 11.97 | 14.11 | 29.00 | 68.00 | 88.40 | 74.62 | 77.94 | 92.10 |
| dangvantuan/vietnamese-document-embedding | | | | | | | | | | | | | | | |
| Dense | 22.70 | 46.90 | 29.62 | 33.71 | 56.20 | 10.70 | 27.40 | 15.02 | 17.90 | 35.70 | 75.70 | 95.60 | 83.00 | 86.10 | 97.40 |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | | | | | | | | | | | | | | | |
| Dense | 4.50 | 14.80 | 7.16 | 8.95 | 20.50 | 4.70 | 16.10 | 7.30 | 9.33 | 22.10 | 55.90 | 81.30 | 64.15 | 68.28 | 87.70 |
| intfloat/multilingual-e5-large | | | | | | | | | | | | | | | |
| Dense | 23.90 | 49.20 | 31.09 | 35.38 | 57.00 | 11.80 | 29.00 | 16.77 | 19.66 | 37.00 | 83.60 | 97.90 | 89.29 | 91.43 | 98.90 |
| + BM25 ($\alpha$) | 29.00 | 57.90 | 37.79 | 42.56 | 66.80 | 12.40 | **31.60** | 17.84 | 21.10 | 39.40 | **89.70** | 98.90 | **93.62** | **94.96** | **99.50** |
| + BM25 (RRF) | 28.40 | 56.90 | 36.69 | 41.48 | 67.80 | 11.90 | 30.20 | 16.94 | 20.07 | 38.30 | 86.50 | 98.70 | 91.27 | 93.12 | 99.40 |
| jinaai/jina-embeddings-v3 | | | | | | | | | | | | | | | |
| Dense | 24.00 | 51.90 | 32.03 | 36.74 | 59.80 | 9.90 | 27.50 | 14.73 | 17.74 | 36.00 | 72.20 | 93.20 | 79.54 | 82.87 | 95.60 |
| + BM25 ($\alpha$) | 30.00 | 61.40 | **39.53** | **44.73** | 68.30 | **13.00** | 30.70 | 17.82 | 20.84 | 40.10 | 84.60 | 98.00 | 89.77 | 91.81 | 98.60 |
| + TF–IDF (RRF) | 27.20 | 60.50 | 36.62 | 42.24 | 69.80 | 12.10 | 28.90 | 16.50 | 19.40 | 37.40 | 74.50 | 97.10 | 83.04 | 86.53 | 98.30 |
| + BM25 (RRF) | **30.70** | 60.60 | 39.41 | 44.43 | **69.90** | 12.70 | 30.50 | 17.10 | 20.20 | 39.20 | 80.40 | 97.60 | 87.26 | 89.84 | 98.90 |
| BAAI/bge-m3 | | | | | | | | | | | | | | | |
| Dense | 24.20 | 51.90 | 32.49 | 37.10 | 59.90 | 12.40 | 30.80 | 17.25 | 20.42 | 40.00 | 80.60 | 96.40 | 86.62 | 89.04 | 98.40 |
| + BM25 ($\alpha$) | 30.50 | 59.20 | 39.20 | 43.97 | 66.10 | 12.40 | 31.10 | 18.02 | 21.12 | **40.80** | 88.30 | 99.00 | 92.41 | 94.04 | 99.30 |
| Snowflake/snowflake-arctic-embed-l-v2.0 | | | | | | | | | | | | | | | |
| Dense | 20.70 | 49.60 | 28.91 | 33.81 | 59.30 | 9.20 | 24.80 | 13.66 | 16.29 | 33.10 | 75.40 | 94.70 | 82.31 | 85.33 | 96.70 |
| + BM25 ($\alpha$) | 29.40 | **61.80** | 38.91 | 44.34 | 69.60 | 11.50 | 28.30 | 16.37 | 19.21 | 36.60 | 84.50 | 97.70 | 89.71 | 91.70 | 98.60 |
| Alibaba-NLP/gte-multilingual-base | | | | | | | | | | | | | | | |
| Dense | 20.20 | 46.70 | 28.12 | 32.53 | 55.00 | 9.00 | 27.00 | 13.74 | 16.83 | 33.90 | 75.00 | 95.20 | 82.34 | 85.49 | 97.00 |
| BAAI/bge-multilingual-gemma2 | | | | | | | | | | | | | | | |
| Dense | 18.80 | 41.20 | 25.26 | 29.03 | 49.00 | 9.20 | 22.10 | 12.64 | 14.86 | 27.60 | 70.90 | 93.00 | 78.74 | 82.23 | 96.00 |
| google/EmbeddingGemma-300m | | | | | | | | | | | | | | | |
| Dense | 21.10 | 47.40 | 28.72 | 33.14 | 55.60 | 10.60 | 27.50 | 15.19 | 18.09 | 34.50 | 76.90 | 94.90 | 83.45 | 86.26 | 97.40 |
| + BM25 ($\alpha$) | 28.20 | 57.90 | 37.03 | 41.98 | 67.30 | 12.50 | 31.10 | 17.35 | 20.57 | 39.10 | 86.20 | 98.50 | 91.24 | 93.06 | 99.10 |
| Alibaba-NLP/gte-Qwen2-1.5B-instruct | | | | | | | | | | | | | | | |
| Dense | 3.60 | 11.20 | 5.40 | 6.75 | 16.40 | 2.10 | 9.80 | 4.07 | 5.40 | 13.50 | 56.70 | 84.10 | 65.64 | 70.09 | 88.50 |
| Qwen/Qwen3-Embedding-0.6B | | | | | | | | | | | | | | | |
| Dense | 23.50 | 48.10 | 30.71 | 34.82 | 55.80 | 11.50 | 28.10 | 15.97 | 18.81 | 35.00 | 73.70 | 94.10 | 80.76 | 84.01 | 97.20 |

**I8.** Zero-shot learned-sparse and late-interaction methods underperform strong lexical baselines without Vietnamese-specific adaptation across domains.

**I9.** Domain difficulty varies markedly within the benchmark, with structured legal and encyclopedic corpora approaching saturation under strong hybrid configurations, while informal and user-generated datasets remain substantially more challenging.

**I10.** Accordingly, meaningful progress in Vietnamese retrieval should be evaluated primarily on the most challenging domains, where performance differences remain discriminative.

### 4.5 Error Analysis

We analyze representative failure cases from the lowest-scoring datasets, including EduCoQA (Education), CSConDa (Customer Support), and VlogQA/ViRe4MRC (Lifestyle & Reviews). These errors reveal recurring linguistic and task-specific challenges that limit retrieval effectiveness; detailed examples are provided in Appendix G.1.

**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 reduce lexical overlap for sparse retrievers and can introduce semantic drift for dense encoders. Spelling variation, typos, and missing diacritics further degrade retrieval quality, with inconsistent casing having similar effects.

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

**E3.** Entity, scope, and intent ambiguity remains unresolved, especially in education and customer-service domains. Underspecified queries such as *"trưởng khoa là ai?" [who is the dean?]* can refer 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 data (pricing, eligibility, compliance) further increase the risk of retrieving passages that match surface terms but miss the user's actual information need.

## 5 Discussion

Our findings have practical implications for Vietnamese QA and cross-lingual retrieval.

**Implications for system design.** Hybrid lexical–semantic retrieval emerges as the default configuration for Vietnamese QA systems. Multilingual encoders offer robustness on conversational and customer-support data, while Vietnamese-specialized models remain competitive on formal, institutionally grounded corpora. These findings suggest that domain-aware retrieval pipelines may be preferable to a single uniform architecture.

**Limitations and open challenges.** A clear performance gap persists between structured and informal datasets. Legal and encyclopedic corpora approach saturation, whereas EduCoQA, CSConDa, VlogQA, and ViRe4MRC remain substantially more difficult. This disparity reflects recurring issues such as noisy queries, register mismatches, and ambiguous user intent. The near-ceiling performance on structured benchmarks also suggests that current evaluation settings may be insufficiently challenging for strong retrievers.

**Future directions.** Future work should focus on constructing harder and more linguistically diverse benchmarks with richer lexical variation and stronger negatives. Domain-adaptive training strategies may further improve robustness on informal corpora. Extending evaluation to multimodal and speech-driven retrieval would better reflect real-world deployment scenarios.

## 6 Conclusion

We presented the first domain-normalized, multi-domain benchmark for Vietnamese information retrieval under a unified and reproducible evaluation protocol, spanning six domains and ten datasets, and introduced two realistic QA datasets derived from authentic educational counseling and customer-service interactions. Through a systematic comparison of lexical, neural-sparse, late-interaction, dense, and hybrid paradigms, we identify several consistent trends. Lexical–semantic hybridization, particularly BM25-based linear fusion, is the most reliable retrieval strategy across domains. Vietnamese-specialized encoders perform strongly on formal institutional corpora but generalize less effectively to conversational and lifestyle-oriented data, whereas multilingual encoders provide stronger cross-domain robustness. We further show that model scale alone is not a dependable predictor of retrieval effectiveness, and that informal, user-generated corpora remain substantially more challenging than structured legal and encyclopedic datasets.

By isolating paradigm-level and architectural effects within a controlled multi-domain setting, our benchmark clarifies trade-offs that were previously obscured by single-domain or curated evaluations. Beyond reporting comparative results, this work establishes a reproducible foundation for Vietnamese retrieval research and provides actionable guidance for designing robust RAG systems in low-resource settings. We hope these resources and findings will enable harder and more realistic benchmarks, stronger domain-adaptive retrieval methods, and more reliable evaluation standards for Vietnamese and other underrepresented languages.

## Limitations

Although our benchmark provides the most comprehensive multi-domain evaluation of Vietnamese retrieval to date, several scope boundaries remain. The study focuses primarily on text-based, single-hop retrieval under a standardized setting of 1,000 query–document pairs per dataset to ensure fair cross-domain comparison; while this design enables controlled analysis, it does not fully capture long-tail query distributions, large-scale corpus effects, or more complex multi-hop and long-context reasoning scenarios. Neural retrievers are evaluated in their original pretrained form under a unified zero-shot protocol to isolate paradigm-level differences, and thus do not reflect potential gains from supervised, instruction-tuned, or domain-adaptive training. In addition, the benchmark centers on Vietnamese text-only corpora, leaving cross-lingual, speech-based, and multimodal retrieval as important directions for future work. Finally, although we introduce two new re-

alistic datasets, informal and highly noisy user-generated content remains comparatively limited, highlighting opportunities for further expansion. These considerations naturally reflect the scope of the present study and point toward promising avenues for advancing robust and inclusive Vietnamese QA systems.

## Supplementary Materials Availability Statement

All datasets used in this study, including previously published resources as well as the newly introduced EduCoQA and CSConDa, together with standardized evaluation subsets and benchmarking scripts, are publicly available at https://github.com/longstnguyen/ViRE. All reused resources comply with their original academic or open-source licenses. EduCoQA and CSConDa are released under the CC BY-NC 4.0 license for non-commercial research use.

## Ethical Considerations

All datasets comply with their original licenses and pose no privacy risks. For the two proposed datasets, all personally identifiable or sensitive information was carefully removed, and only non-sensitive portions are released for research use. The collection and cleaning process, described in Appendix D, followed strict ethical standards under the ACL Code of Ethics.

## 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).

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, 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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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.

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).

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

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.

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.

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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and Léonard Hussenot. 2025. Gemma 3 Technical Report.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving Open Language Models at a Practical Size.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

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).

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 Technical Report.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 39 others. 2024. Qwen2 Technical Report.

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.

## A Experimental Setup

All experiments were conducted on a single NVIDIA A100 GPU with 40GB VRAM. Open-source dense encoders were evaluated using their official Hugging Face[4] checkpoints, while the proprietary OpenAI model was accessed via API. For dense retrieval, we used FAISS (Johnson et al., 2019) for vector indexing and similarity search. The key hyperparameter settings for BM25 and hybrid retrieval methods are summarized in Table 5.

Table 5: Hyperparameter settings for retrieval models.

| Parameter | Value |
|---|---|
| BM25 $k_1$ | 1.5 |
| BM25 $b$ | 0.75 |
| Weighted fusion coefficient ($\alpha$) | 0.7 |
| RRF constant ($c$) | 60 |

## B Model Analysis

Table 6 summarizes the architectural characteristics of the dense embedding models evaluated across our ten Vietnamese retrieval benchmarks. The models span six backbone families: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020), GTE (Li et al., 2023), Gemma (Team et al., 2024, 2025), and Qwen (Yang et al., 2024, 2025). They vary widely in parameter count (118 M–9.2 B) and embedding dimensionality (384–3,584), providing a diverse architectural spectrum for analyzing how model design relates to multi-domain retrieval behavior.

**Model scale and dimensionality are weak predictors.** Spearman correlations between mean nDCG@10 and model size reveal no significant association with either log-parameter count ($\rho = 0.16$, $p = 0.60$) or embedding dimensionality ($\rho = 0.23$, $p = 0.45$). For example, the largest model, bge-multilingual-gemma2 (9.2 B parameters), ranks tenth overall with a mean nDCG@10 of 54.4%, whereas the much smaller bge-m3 (568 M) achieves the highest score at 64.5%. A weak but non-significant correlation between parameter count and cross-domain variance ($\rho = 0.44$, $p = 0.14$) further suggests that larger models may exhibit slightly greater domain sensitivity.

**Architecture family shows limited influence.** A Kruskal–Wallis test across backbone families finds no statistically significant difference in retrieval effectiveness ($H = 7.41$, $p = 0.19$), indicating that architecture alone does not reliably determine performance. Pairwise Mann–Whitney tests identify only one significant difference: XLM-RoBERTa models outperform BERT-based models ($p = 0.030$). At the aggregate level, XLM-RoBERTa achieves the highest average nDCG@10 (63.5%), followed by GTE (61.0%), Gemma (58.4%), RoBERTa (53.9%), Qwen (53.7%), and BERT (42.0%). However, within-family variability remains substantial; for instance, gte-Qwen2-1.5B ranges from 91.77% on ZaloLegalQA to below 6% on several informal datasets.

**Pretraining alignment outweighs raw scale.** Across datasets, multilingual encoders demonstrate the most consistent cross-domain behavior. Models grounded in broad multilingual pretraining, particularly XLM-RoBERTa and GTE encoders such as bge-m3, jina-embeddings-v3, multilingual-e5-large, and Vietnamese_Embedding_v2, tend to achieve stronger average retrieval performance than architectures optimized primarily for scale. This pattern suggests that representational alignment with diverse linguistic contexts is more influential than parameter size alone.

[4] https://huggingface.co/

Table 6: Architectural families, parameter scale, and embedding dimensionality of the evaluated dense retrieval models. The models span six backbone families and vary substantially in size and representational capacity. Models highlighted in red are specifically trained for the Vietnamese language.

| Model ID | Architecture | #Params | Embedding Dim. |
|---|---|---|---|
| OpenAI's text-embedding-3-large | – | – | 3,072 |
| AITeamVN/Vietnamese_Embedding_v2 | XLM-RoBERTa | 567,754,752 | 1,024 |
| bkai-foundation-models/vietnamese-bi-encoder | RoBERTa | 134,998,272 | 768 |
| dangvantuan/vietnamese-document-embedding | GTE | 305,368,320 | 768 |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | BERT | 117,653,760 | 384 |
| intfloat/multilingual-e5-large | XLM-RoBERTa | 559,890,432 | 1,024 |
| jinaai/jina-embeddings-v3 | XLM-RoBERTa | 567,754,752 | 1,024 |
| BAAI/bge-m3 | XLM-RoBERTa | 567,754,752 | 1,024 |
| Snowflake/snowflake-arctic-embed-l-v2.0 | XLM-RoBERTa | 567,754,752 | 1,024 |
| Alibaba-NLP/gte-multilingual-base | GTE | 305,368,320 | 768 |
| BAAI/bge-multilingual-gemma2 | Gemma 2 | 9,241,713,152 | 3,584 |
| google/EmbeddingGemma-300m | Gemma 3 | 302,863,104 | 768 |
| Alibaba-NLP/gte-Qwen2-1.5B-instruct | Qwen 2 | 1,543,268,864 | 1,536 |
| Qwen/Qwen3-Embedding-0.6B | Qwen 3 | 595,776,512 | 1,024 |
| colbert-ir/colbertv2.0 | BERT | 109,482,240 | 768 |
| naver/splade-v3 | BERT | 109,482,240 | 768 |

**Language-specific training offers limited structural advantage.** We compare Vietnamese-specific encoders with multilingual models. Mann–Whitney tests reveal no statistically significant performance difference on Vietnamese regulatory corpora (ALQAC, ZaloLegalQA, ViRHE4QA; $p = 0.79$) or encyclopedic QA (UIT-ViQuAD; $p = 0.66$). Cross-domain variance is also nearly identical between the two groups ($\sigma_{\text{VI}} = 27.23$ vs. $\sigma_{\text{ML}} = 27.09$, $p = 0.29$), and the frequency of top-performing systems across datasets is comparable.

Taken together, these observations suggest that retrieval effectiveness depends less on model scale or language-specific pretraining than on alignment between pretraining coverage and the linguistic characteristics of the target corpus. Consistent with the domain-level analysis, corpus register and discourse structure emerge as stronger determinants of retrieval behavior than architectural provenance.

## C Data Analysis

We provide a quantitative characterization of the datasets in our benchmark. Table 7 reports the number of records, unique contexts, and token-level statistics for queries and contexts. Tokenization is performed using the Vietnamese segmenter from the widely used Underthesea toolkit[5], which provides reliable word segmentation for Vietnamese.

Substantial variation is observed in both query and context lengths across datasets. For instance, VlogQA contains exceptionally long contexts with an average exceeding 2,200 tokens, whereas ViRe4MRC averages fewer than 100 tokens per context. In contrast, query lengths remain consistently short across domains, with means ranging from approximately 9 to 19 tokens.

Crucially, retrieval difficulty does not scale monotonically with context length. Datasets with very long contexts such as VlogQA remain challenging, yet shorter and more structurally homogeneous corpora, particularly in the legal domain, approach performance saturation under strong hybrid configurations. This suggests that linguistic variability and domain structure play a more decisive role in retrieval complexity than raw document length alone.

## D Proposed Datasets

To support realistic and domain-aware evaluation of Vietnamese retrieval, we introduce two new QA datasets with explicit document alignment: ***Edu**cational **C**ounseling **QA*** (EduCoQA) and ***C**ustomer **S**upport **Con**versations **Da**taset* (CSConDa). Both resources consist of naturally occurring user questions paired with authoritative supporting documents, enabling controlled retrieval evaluation while preserving the linguistic variability of real-world Vietnamese.

### D.1 CSConDa

**Data Collection and Alignment.** CSConDa was developed in collaboration with a national provider

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

Table 7: 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 |

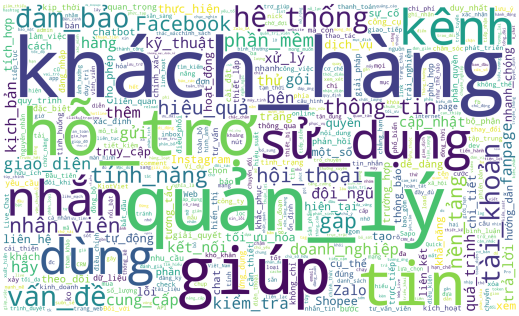


Figure 2: Word cloud of CSConDa highlighting frequent customer-service topics.


Figure 3: Word cloud of EduCoQA illustrating frequent admission-related topics.

of multi-channel customer service solutions. The construction process followed three stages. First, raw chat logs were collected from platforms including Facebook[6], Zalo[7], Shopee[8], and related support channels. Conversations were filtered to retain coherent, task-oriented exchanges while excluding sensitive or inappropriate content. Second, conversations were segmented into contextually consistent QA pairs and anonymized to remove personally identifiable information and system artifacts. Finally, each question was aligned with supporting passages drawn from official help-center articles, service manuals, and policy documents. Segmentation ensured that each aligned document contains sufficient evidence to answer the query, making the dataset directly suitable for retrieval benchmarking.

**Topical Coverage.** CSConDa captures diverse customer-support intents, including pricing inquiries, subscription eligibility, technical troubleshooting, account management, and policy compliance. As illustrated by the word distribution in Figure 2, frequent terms such as *"khách hàng" [customer]*, *"hỗ trợ" [support]*, *"tài khoản" [account]*, and *"quản lý" [management]* reflect its operational and service-oriented focus.

### D.2 EduCoQA

**Data Collection and Alignment.** EduCoQA was curated from authentic university admission counseling activities. To capture natural information needs, we deployed a lightweight RAG-based chatbot at educational fairs and institutional advising events, collecting spontaneous questions from high-school students and parents. Queries were logged from both in-person and online counseling sessions, reflecting genuine concerns regarding admissions, academic programs, and institutional policies. Each query was subsequently aligned with authoritative institutional sources, including official regulations, program descriptions, and departmental materials. Advisor responses were incorpo-

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

rated to ensure factual consistency, and supporting passages were segmented to provide precise and verifiable evidence for retrieval evaluation.

**Topical Coverage.** EduCoQA encompasses university information, admission criteria, scholarships, curriculum structures, and career pathways. The dominant themes, visualized in Figure 3, include *"tuyển sinh" [admissions]*, *"chương trình" [program]*, *"bộ môn" [department]*, *"sinh viên" [student]*, and *"kỹ thuật" [engineering]*, reflecting its academic counseling orientation.

### D.3 Ethical Considerations

All data were anonymized prior to release, and personally identifiable information was removed during preprocessing. Customer-service conversations were filtered to exclude sensitive content, and educational documents originate from publicly accessible institutional sources. Both datasets are released exclusively for research use under terms that respect user privacy and institutional integrity. All contributors involved in dataset construction signed Non-Disclosure Agreements to ensure confidentiality and compliance with ethical standards.

### D.4 Examples

Representative examples from both datasets are provided in Appendix G.1, where we also present a qualitative failure analysis. These cases highlight the linguistic variability, informality, implicit intent, and ambiguity present in real-world Vietnamese queries, illustrating why EduCoQA and CSConDa constitute challenging benchmarks for retrieval models.

## E Detailed Analysis

We provide domain-level analyses that contextualize the quantitative trends summarized in Section 4.4. The first subsection characterizes structural retrieval regimes across domains, while the second introduces a rank-based aggregation framework to assess cross-domain consistency.

### E.1 Domain-Level Analysis

**Education.** The education domain exhibits a clear bifurcation between informal counseling queries (EduCoQA) and structured regulatory text (ViRHE4QA). On EduCoQA, even the strongest hybrid configuration reaches only MRR@10 $\approx$ 35, reflecting paraphrastic and colloquial student queries with limited lexical overlap, as reported in Table 12. This difficulty is amplified by systematic cross-register mismatch: queries follow informal conversational patterns, whereas relevant passages are drawn from formal institutional documents, widening the vocabulary gap that lexical matching cannot reliably bridge. Multilingual encoders such as multilingual-e5-large perform best, suggesting that broader semantic coverage is essential for informal academic counseling language. In contrast, ViRHE4QA achieves substantially higher effectiveness under lexical–dense hybrids (MRR@10 $>$ 80), with BM25 alone already highly competitive. Here, queries and passages share the same formal regulatory register, enabling strong term-level alignment without reliance on deeper semantic abstraction. Vocabulary consistency and regulatory formality therefore dominate ranking behavior, with Vietnamese-specific pretraining providing an additional advantage.

**Customer Support.** CSConDa represents a short-query, intent-heavy retrieval regime, with detailed per-method scores provided in Table 13. Dense encoders benefit from conversational pretraining, and hybridization yields consistent improvements, yet absolute performance remains moderate (best MRR@10 $\approx$ 44). Performance limitations arise from paraphrastic phrasing and sparse lexical anchors, which weaken both exact term matching and semantic stability. A particularly challenging structural characteristic is cross-brand reference: queries often mention competitor products while the relevant passage describes a different system. Neither surface-level lexical overlap nor naive embedding similarity reliably resolves this intent-level indirection, indicating that successful retrieval may require external product knowledge. Compared with formal domains, improvements in this regime remain incremental rather than transformative.

**Legal.** Legal corpora exhibit highly regularized terminology and citation patterns, with comprehensive results summarized in Table 14. BM25 alone yields strong performance, and hybrid methods push MRR@10 beyond 95 on ALQAC. Notably, BM25 standalone achieves MRR@10 $\approx$ 92 on ALQAC, the highest single-method lexical score observed in the benchmark, reflecting the exceptional formulaic consistency of Vietnamese statutory text. Performance gaps among encoders narrow considerably, indicating that standardized legislative vocabulary allows lexical sig-

nals to dominate ranking behavior. On more naturally phrased legal queries such as ZaloLegalQA, Vietnamese-aligned encoders regain a modest advantage, though overall headroom remains limited.

**Healthcare.** Healthcare reveals two distinct retrieval patterns, with quantitative comparisons presented in Table 15. In medical news (ViNewsQA), lexical matching remains competitive, whereas in clinically oriented QA (ViMedAQA), dense encoders substantially outperform BM25 from the outset. This divergence stems from corpus structure: ViNewsQA passages employ accessible journalistic language, while ViMedAQA passages contain specialized clinical terminology whose semantic relationships are not recoverable through unigram overlap alone. Hybrid fusion benefits both settings, but for different reasons, as journalistic content rewards lexical alignment whereas clinical discourse demands stronger semantic modeling. The contrast illustrates how structural characteristics of the corpus, rather than topical label alone, determine the dominant retrieval signal.

**Lifestyle & Reviews.** Informal lifestyle corpora constitute the most challenging regime in the benchmark, as reflected in the detailed results of Table 16. Conversational vlog transcripts reward models capable of capturing extended discourse context, yet even the strongest hybrids reach only MRR@10 $\approx$ 39. VlogQA passages are raw ASR transcriptions lacking sentence boundaries, punctuation, and standardized orthography, producing long spans of unstructured text that current retrieval architectures are not optimized to process. Product and food reviews (ViRe4MRC) are harder still, with best MRR@10 $\approx$ 18, reflecting fragmented and opinion-driven snippets that provide few stable lexical anchors. These properties expose the largest robustness gap in Vietnamese retrieval.

**Cross-domain Open Knowledge.** On UIT-ViQuAD, full evaluation results are reported in Table 17. BM25 provides a strong baseline due to high lexical overlap in encyclopedic text, while multilingual hybrids achieve near-maximal effectiveness (MRR@10 $\approx$ 94, R@20 $>$ 99%). The dense lexical anchor density results from the extractive QA construction, where answer spans are verbatim substrings of the source passage, naturally inducing term overlap between query and document. Performance differences are therefore narrow, and ranking variance correspondingly limited.

**Cross-domain synthesis.** Across domains, retrieval effectiveness correlates more strongly with linguistic regularity and discourse structure than with topical category. Structured and vocabulary-consistent corpora tend to exhibit lexical dominance and limited ranking dispersion, whereas informal, noisy, or fragmentary corpora expose the limitations of current dense and hybrid retrieval approaches. This structural perspective helps explain why performance variation across domains often exceeds variation across model architectures.

### E.2 Rank-Based Evaluation

To enable fair and interpretable comparison, we employ a rank-based aggregation scheme over the set of benchmark datasets $\mathcal{B}$. This framework highlights not only the strongest methods within each dataset, but also the most consistent performers across heterogeneous domains.

**Per-metric ranking.** Let $\mathcal{K}$ denote the set of evaluation metrics (e.g., P@1, R@10, MRR@10, nDCG@10, R@20). For each dataset $b \in \mathcal{B}$, metric $m \in \mathcal{K}$, and retrieval method $s \in \mathcal{S}$, we define the evaluation score as

$$V_{b,m}(s) = \mathcal{M}_m^b(s), \tag{14}$$

where $\mathcal{M}_m^b(\cdot)$ denotes the metric-specific evaluation function. The corresponding rank is

$$r_{b,m}(s) = 1 + \left|\{\, s' \in \mathcal{S} : V_{b,m}(s') > V_{b,m}(s) \,\}\right|, \tag{15}$$

with ties assigned the smallest rank in their group.

**Normalized rank.** To account for differing numbers of evaluated systems per dataset, ranks are normalized into $[0, 1]$:

$$\tilde{r}_{b,m}(s) = \frac{r_{b,m}(s) - 1}{|\mathcal{S}_b| - 1}, \tag{16}$$

where $|\mathcal{S}_b|$ is the number of systems evaluated on dataset $b$. Here, $\tilde{r}_{b,m}(s) = 0$ denotes best performance and $\tilde{r}_{b,m}(s) = 1$ worst.

**Dataset-level aggregation.** For each dataset $b$, the aggregate rank of method $s$ is

$$R_b(s) = \sum_{m \in \mathcal{K}} \tilde{r}_{b,m}(s). \tag{17}$$

**Domain and overall aggregation.** For domain $g \in \mathcal{G}$, let $\mathcal{B}^{(g)} \subseteq \mathcal{B}$ denote the corresponding dataset subset. The domain-level score is

$$R_{\mathcal{B}^{(g)}}(s) = \sum_{b \in \mathcal{B}^{(g)}} R_b(s), \tag{18}$$

Table 8: Top–3 retrieval methods per domain and overall under rank-based evaluation. All hybrid systems use linear interpolation with BM25 or TF–IDF unless otherwise specified.

| Domain | 1st Place | 2nd Place | 3rd Place |
|---|---|---|---|
| Education | multilingual-e5-large + BM25 | bge-m3 + BM25 | snowflake-arctic-embed-l-v2.0 + BM25 |
| Customer Support | jina-embeddings-v3 + BM25 | text-embedding-3-large + BM25 | jina-embeddings-v3 + TF–IDF |
| Legal | Vietnamese_Embedding_v2 + TF–IDF | Vietnamese_Embedding_v2 + BM25 | text-embedding-3-large + BM25 |
| Healthcare | multilingual-e5-large + BM25 | EmbeddingGemma-300m + BM25 | text-embedding-3-large + BM25 |
| Lifestyle & Reviews | jina-embeddings-v3 + BM25 | bge-m3 + BM25 | jina-embeddings-v3 + BM25 (RRF) |
| Cross-domain | multilingual-e5-large + BM25 | Vietnamese_Embedding_v2 + BM25 | bge-m3 + BM25 |
| Overall | multilingual-e5-large + BM25 | bge-m3 + BM25 | jina-embeddings-v3 + BM25 |

and the overall benchmark score is

$$R(s) = \sum_{g \in \mathcal{G}} R_{\mathcal{B}^{(g)}}(s), \qquad \bar{R}(s) = \frac{R(s)}{|\mathcal{B}|}. \quad (19)$$

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

**Results.** Table 8 summarizes the top–3 methods under the proposed aggregation scheme. Hybrid systems based on linear interpolation with BM25 consistently occupy leading positions across domains, confirming lexical–semantic integration as the most stable retrieval strategy for Vietnamese text. The configuration built upon multilingual-e5-large with BM25 interpolation achieves the strongest overall average rank, indicating robust cross-domain generalization. Models derived from jina-embeddings-v3 perform particularly well in conversational regimes such as Customer Support and Lifestyle & Reviews, whereas Vietnamese_Embedding_v2 demonstrates clear specialization in legal settings. Overall, the rank-based perspective suggests that cross-regime stability, rather than peak performance on a single dataset, defines benchmark-level retrieval strength.

# F Supplementary Experiments

## F.1 Subset Evaluation Validity

As described in Section 4.2, the main benchmark evaluates each dataset on a randomly sampled subset of 1,000 query–document pairs in order to standardize evaluation size across corpora and retrieval configurations. A natural concern is whether such sampling could alter model rankings or mask patterns that emerge only at full-corpus scale. To assess this possibility, we run the complete retrieval pipeline on the full UIT-ViQuAD corpus and compare the results with those obtained from its corresponding 1,000-pair subset using the strongest dense retrieval backbones.

Table 9 reports retrieval results for 25 configurations, covering five dense encoders and five retrieval variants per encoder. The subset evaluation closely reproduces the ranking behavior observed on the full corpus. Spearman correlation between full-corpus and subset nDCG@10 scores reaches $\rho = 0.899$ ($p < 10^{-9}$), while Kendall's $\tau = 0.773$ ($p < 10^{-7}$), indicating strong agreement between the two evaluation regimes.

The highest-performing configuration is identical in both settings, namely multilingual-e5-large combined with BM25 using linear interpolation. The three best-performing configurations are likewise preserved across the two regimes. Although subset scores are consistently higher—by approximately 10.1 percentage points on average due to the smaller document pool—this shift affects all configurations in a similar manner and therefore does not alter their relative ordering.

Overall, the analysis confirms that the 1,000-pair subset provides a reliable approximation of full-corpus evaluation. While absolute metric values increase under the smaller candidate pool, the relative behavior of retrieval methods and the main conclusions of the benchmark remain unchanged.

## F.2 Sensitivity to the Interpolation Parameter

To further examine lexical–semantic balance in hybrid retrieval, we conduct an additional experiment on the full corpus of UIT-ViQuAD. We evaluate the best-performing configuration that combines BM25 with the dense encoder multilingual-e5-large, while varying the interpolation parameter $\alpha$ from 0 to 1 with a step size of 0.05.

Figure 4 illustrates the impact of $\alpha$ on retrieval performance. All metrics improve as $\alpha$ increases from 0 to around 0.6, indicating that dense semantic similarity contributes strongly to effective ranking. Performance peaks within $\alpha \in [0.6, 0.8]$, after which it gradually declines as the lexical contribution from BM25 becomes too weak.

Table 9: Retrieval results on UIT-ViQuAD under two evaluation settings.

| Method | Full corpus | | | | | Subset (1,000 query–document) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| text-embedding-3-large | | | | | | | | | | |
| Dense | 55.00 | 84.00 | 65.00 | 70.00 | 89.00 | 71.20 | 92.40 | 78.75 | 82.09 | 96.00 |
| + TF–IDF ($\alpha$) | 65.00 | 91.00 | 74.00 | 78.00 | 95.00 | 79.30 | 97.20 | 86.24 | 88.96 | 98.20 |
| + BM25 ($\alpha$) | 70.00 | 93.00 | 78.00 | 82.00 | 96.00 | 82.90 | 97.00 | 88.46 | 90.58 | 98.70 |
| + TF–IDF (RRF) | 58.00 | 90.00 | 69.00 | 74.00 | 95.00 | 74.10 | 96.10 | 82.18 | 85.60 | 98.30 |
| + BM25 (RRF) | 66.00 | 92.00 | 75.00 | 79.00 | 96.00 | 77.80 | 95.80 | 84.59 | 87.37 | 98.70 |
| AITeamVN/Vietnamese_Embedding_v2 | | | | | | | | | | |
| Dense | 66.00 | 92.00 | 75.00 | 79.00 | 96.00 | 82.20 | 98.10 | 88.24 | 90.67 | 99.00 |
| + TF–IDF ($\alpha$) | 70.00 | 95.00 | 79.00 | 83.00 | 97.00 | 84.40 | **99.30** | 90.18 | 92.45 | 99.30 |
| + BM25 ($\alpha$) | 77.00 | 96.00 | 84.00 | 87.00 | 98.00 | 89.10 | 99.20 | 93.00 | 94.53 | 99.30 |
| + TF–IDF (RRF) | 65.00 | 93.00 | 74.00 | 79.00 | 97.00 | 79.00 | 97.60 | 86.02 | 88.87 | 99.30 |
| + BM25 (RRF) | 73.00 | 95.00 | 81.00 | 84.00 | 97.00 | 82.10 | 97.50 | 87.95 | 90.32 | 99.10 |
| intfloat/multilingual-e5-large | | | | | | | | | | |
| Dense | 70.00 | 93.00 | 78.00 | 82.00 | 96.00 | 83.60 | 97.90 | 89.29 | 91.43 | 98.90 |
| + TF–IDF ($\alpha$) | 70.00 | 95.00 | 79.00 | 83.00 | 97.00 | 85.70 | 99.00 | 90.83 | 92.86 | 99.30 |
| + BM25 ($\alpha$) | **77.00** | **96.00** | **84.00** | **87.00** | **98.00** | **89.70** | 98.90 | **93.62** | **94.96** | **99.50** |
| + TF–IDF (RRF) | 64.00 | 93.00 | 74.00 | 78.00 | 97.00 | 78.60 | 98.40 | 86.52 | 89.49 | 99.20 |
| + BM25 (RRF) | 73.00 | 95.00 | 81.00 | 84.00 | 97.00 | 86.50 | 98.70 | 91.27 | 93.12 | 99.40 |
| jinaai/jina-embeddings-v3 | | | | | | | | | | |
| Dense | 55.00 | 84.00 | 65.00 | 69.00 | 89.00 | 72.20 | 93.20 | 79.54 | 82.87 | 95.60 |
| + TF–IDF ($\alpha$) | 63.00 | 91.00 | 73.00 | 77.00 | 95.00 | 79.70 | 97.80 | 86.36 | 89.18 | 98.40 |
| + BM25 ($\alpha$) | 71.00 | 93.00 | 79.00 | 82.00 | 96.00 | 84.60 | 98.00 | 89.77 | 91.81 | 98.60 |
| + TF–IDF (RRF) | 57.00 | 89.00 | 68.00 | 73.00 | 94.00 | 74.50 | 97.10 | 83.04 | 86.53 | 98.30 |
| + BM25 (RRF) | 66.00 | 92.00 | 75.00 | 79.00 | 96.00 | 80.40 | 97.60 | 87.26 | 89.84 | 98.90 |
| BAAI/bge-m3 | | | | | | | | | | |
| Dense | 66.00 | 91.00 | 75.00 | 79.00 | 94.00 | 80.60 | 96.40 | 86.62 | 89.04 | 98.40 |
| + TF–IDF ($\alpha$) | 68.00 | 94.00 | 77.00 | 81.00 | 97.00 | 83.20 | 98.70 | 89.48 | 91.78 | 99.20 |
| + BM25 ($\alpha$) | 75.00 | 96.00 | 83.00 | 86.00 | 98.00 | 88.30 | 99.00 | 92.41 | 94.04 | 99.30 |
| + TF–IDF (RRF) | 62.00 | 92.00 | 72.00 | 77.00 | 96.00 | 77.20 | 97.60 | 84.95 | 88.07 | 99.00 |
| + BM25 (RRF) | 71.00 | 94.00 | 79.00 | 83.00 | 97.00 | 80.80 | 97.30 | 87.15 | 89.67 | 99.00 |

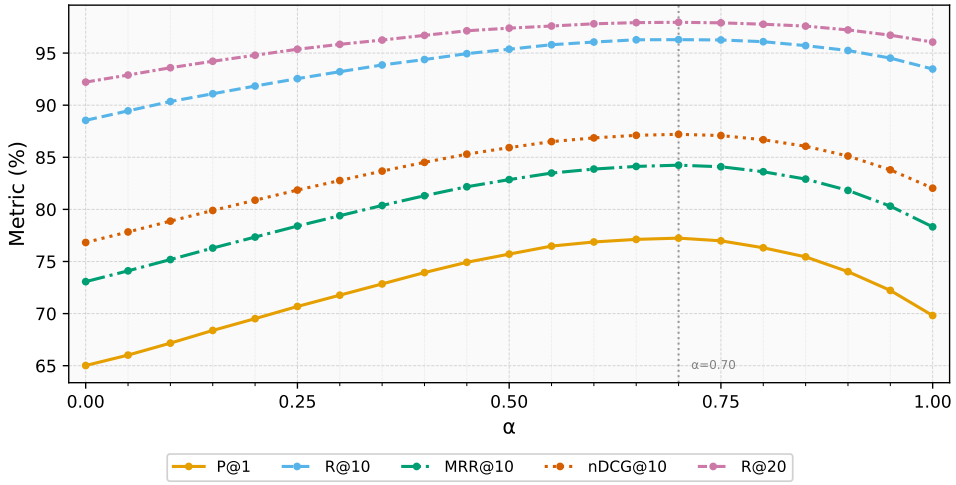


Figure 4: Effect of the interpolation parameter $\alpha$ in hybrid retrieval on the UIT-ViQuAD dataset.

These results highlight the complementary roles of lexical and semantic retrieval. Dense representations capture semantic similarity, whereas BM25 preserves exact lexical signals such as rare terms and named entities. Balancing these signals improves ranking stability by reducing semantic drift while maintaining lexical grounding. We therefore adopt $\alpha = 0.7$ for all hybrid experiments in the main benchmark, as detailed in Appendix A.

## G Qualitative Analysis

### G.1 Representative Error Cases

Table 10 presents representative retrieval failures discussed in Section 4.5, illustrating error types (E1–E3) with real samples drawn from the lowest-performing datasets. Each case contains the original Vietnamese query, its English translation, the corresponding gold passage, and a concise diagnostic note. All examples reflect failure patterns consistently observed across all retrieval methods within each dataset.

### G.2 Representative Successful Retrieval

We complement the error analysis with representative success cases from the highest-performing benchmarks, including ViRHE4QA (Education), ALQAC and ZaloLegalQA (Legal), ViNewsQA and ViMedAQA (Healthcare), and UIT-ViQuAD (Open-domain). These datasets share structural properties that make relevance signals explicit and stable for both sparse and dense retrievers; examples are shown in Table 11. Quantitatively, the best-performing method on each benchmark achieves an average nDCG@10 of 89.9% on these six benchmarks versus 38.7% on the four most challenging benchmarks (EduCoQA, CSConDa, VlogQA, ViRe4MRC), a 51.2-point gap.

**S1.** Near-verbatim lexical overlap provides a dominant retrieval cue in structured and definition-oriented corpora. In statutory benchmarks, offense predicates such as *"chiếm đoạt di vật của tử sĩ" [appropriating the relics of a fallen soldier]* are repeated in both the article heading and the sentencing clause, enabling BM25 to rank the correct provision with minimal ambiguity. A similar pattern appears in clinical QA, where phrases such as *"bệnh thận giai đoạn cuối" [end-stage renal disease]* and *"lọc máu thường xuyên" [regular dialysis]* recur verbatim between the question and definition-style evidence, yielding strong term matching and consistent with a ceiling effect.

**S2.** Register consistency further stabilizes retrieval by reducing stylistic mismatch. Institutional and legal benchmarks often follow standardized administrative phrasing and domain-specific terminology (e.g., *"Công an xã" [commune-level police]*, *"xử phạt" [to impose a fine / sanction]*, and abbreviation-heavy glossary entries), so surface cues transfer directly across the query–document boundary. Success can also arise from intent-level alignment: questions framed as *"có đúng không" [is it lawful]* are often addressed by passages describing authority, jurisdiction, or procedural constraints, even when the query contains additional situational details.

**S3.** Extractive QA collection preserves lexical anchors by construction. In article-grounded QA, queries such as *"hợp chất trong vỏ nho" [a compound in grape skins]* closely track the source sentence describing *Resveratrol*; in Wikipedia-style QA, fragments like *"tính cách đáng sợ" [fearsome character]* and *"người đương thời" [contemporaries]* are often carried over from the evidence sentence into the question. This constrains the candidate space and reduces paraphrastic or cross-register inference.

Overall, these benchmarks are "easy" not because Vietnamese retrieval is solved, but because query–passage alignment is structurally high: terminology is repeated, register is controlled, and many questions preserve evidence phrasing. As a result, architectural differences are less discriminative here and become more pronounced on noisier conversational and user-generated domains.

Table 10: 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.)<br>**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.)<br>→ 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?)<br>**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].)<br>→ 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?)<br>**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.)<br>→ 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?)<br>**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...)<br>→ 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?)<br>**Gold passage:** k chề vào đâu đc đáp ứng tất cả các nhu cầu [...] vô cùng mượt mà [...] ☻☻☻ (**English:** No complaints at all, meets all needs [...] super smooth performance [...] ☻.)<br>→ 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. |

Table 11: Representative successful retrieval examples from high-performing datasets. Only key passage fragments are shown; omitted content is denoted by "[...]".

| Dataset | Example |
|---|---|
| **ViRHE4QA** | **Query:** Hội đồng bảo vệ KLTN còn được gọi tắt là gì? (**English:** What is the abbreviation for the Thesis Defense Committee?)<br>**Gold passage:** Điều 2. Một số thuật ngữ, chữ viết tắt sử dụng trong quy định này\nKLTN: Khóa luận tốt nghiệp\nĐHCNTT: Đại học Công nghệ Thông tin – ĐHQG-HCM\nTrường: Trường ĐHCNTT\nP.ĐTĐH: Phòng Đào tạo Đại học\nKhoa: gọi chung cho Khoa, Bộ môn quản lý sinh viên\nHội đồng: Hội đồng bảo vệ KLTN [...] (**English:** Article 2. Terms and abbreviations used in this regulation\nKLTN: Graduation Thesis\nĐHCNTT: University of Information Technology – VNU-HCM\nTrường: UIT\nP.ĐTĐH: Undergraduate Training Office\nKhoa: general term for the Faculty/Department managing students\nHội đồng: Thesis Defense Committee [...]) |
| **ALQAC** | **Query:** Chiếm đoạt di vật của tử sĩ có thể bị phạt tù lên đến bao nhiêu năm? (**English:** What is the maximum prison sentence for appropriating the relics of a fallen soldier?)<br>**Gold passage:** Tội chiếm đoạt hoặc hủy hoại di vật của tử sỹ\n\n1. Người nào chiếm đoạt hoặc hủy hoại di vật của tử sỹ, thì bị phạt cải tạo không giam giữ đến 03 năm hoặc phạt tù từ 06 tháng đến 03 năm.\n\n2. Phạm tội thuộc một trong các trường hợp sau đây, thì bị phạt tù từ 02 năm đến 07 năm:\n\na) Là chỉ huy hoặc sĩ quan;\n\nb) Chiếm đoạt hoặc hủy hoại di vật của 02 tử sỹ trở lên. (**English:** Offense of appropriating or destroying the relics of a fallen soldier\n\n1. Any person who appropriates or destroys the relics of a fallen soldier is subject to non-custodial reform for up to 3 years or imprisonment from 6 months to 3 years.\n\n2. Aggravated cases carrying 2–7 years of imprisonment include:\n\na) the offender is a commander or officer;\n\nb) the relics of two or more fallen soldiers are appropriated or destroyed.) |
| **ZaloLegalQA** | **Query:** Công an xã xử phạt lỗi không mang bằng lái xe có đúng không? (**English:** Is it lawful for commune-level police to fine drivers for not carrying a driving licence?)<br>**Gold passage:** 1. Bố trí lực lượng tham gia tuần tra, kiểm soát trật tự, an toàn giao thông theo kế hoạch.\n2. Thống kê, báo cáo các vụ vi phạm pháp luật, tai nạn giao thông đường bộ [...]\n4. Lực lượng Công an xã chỉ được tuần tra, kiểm soát trên các tuyến đường liên xã, liên thôn [...] và xử lý các hành vi vi phạm trật tự, an toàn giao thông sau: điều khiển xe mô tô, xe gắn máy không đội mũ bảo hiểm, chở quá số người quy định [...] Nghiêm cấm việc Công an xã dừng xe, kiểm soát trên các tuyến quốc lộ, tỉnh lộ. (**English:** 1. Deploy forces to patrol and control traffic order and safety as planned. [...]\n4. Commune-level police may only patrol inter-commune roads within their jurisdiction and may only sanction the following violations: riding motorcycles without helmets, carrying more passengers than permitted [...] Commune-level police are strictly prohibited from stopping vehicles on national or provincial highways.) |
| **ViNewsQA** | **Query:** Chất bổ trong vỏ nho có tác dụng gì? (**English:** What are the benefits of the compound found in grape skins?)<br>**Gold passage:** Resveratrol là một hợp chất trong vỏ nho có khả năng chống oxy hóa, chống nấm mốc và ký sinh trùng. [...] (**English:** Resveratrol is a compound found in grape skins with antioxidant, antifungal, and antiparasitic properties. [...]) |
| **ViMedAQA** | **Query:** Những người mắc bệnh thận giai đoạn cuối có cần lọc máu thường xuyên không? (**English:** Do patients with end-stage renal disease require regular dialysis to survive?)<br>**Gold passage:** Mất hoàn toàn chức năng của thận, thường là do bệnh thận mạn tiến triển. Những người mắc bệnh thận giai đoạn cuối cần yêu cầu lọc máu thường xuyên để sống sót. (**English:** Complete loss of kidney function, usually due to progressive chronic kidney disease. Patients with end-stage renal disease require regular dialysis to survive.) |
| **UIT-ViQuAD** | **Query:** Mặc cho tính cách đáng sợ của mình, Edward vẫn được người đương thời đánh giá như thế nào đối với việc trị vì đất nước? (**English:** Despite his fearsome character, how did Edward's contemporaries assess his reign?)<br>**Gold passage:** Tuy nhiên mặc dù có những tính cách đáng sợ như vậy, người cùng thời với Edward coi ông là một vị vua có năng lực, thậm chí là một vị vua lý tưởng. Dù không được thần dân yêu thương, ông vẫn nhận được sự kính sợ và tôn trọng. [...] (**English:** However, despite such fearsome qualities, Edward's contemporaries considered him a capable ruler, even an ideal king. Though not loved by his subjects, he still received their fear and respect. [...]) |

Table 12: Retrieval results on **Education Domain** (EduCoQA and ViRHE4QA). Best in **bold**, second-best underlined.

| Method | EduCoQA | | | | | ViRHE4QA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| TF–IDF | 14.68 | 42.47 | 22.23 | 26.99 | 53.82 | 55.60 | 92.00 | 67.70 | 73.60 | 95.90 |
| BM25 | 14.68 | 43.44 | 23.09 | 27.93 | 53.42 | 65.80 | 93.50 | 76.05 | 80.34 | 96.90 |
| ColBERT | 14.48 | 39.33 | 21.54 | 25.75 | 48.73 | 42.90 | 73.40 | 52.49 | 57.50 | 81.80 |
| SPLADE | 11.35 | 30.53 | 16.62 | 19.91 | 39.92 | 35.40 | 72.90 | 46.57 | 52.83 | 82.20 |
| text-embedding-3-large | | | | | | | | | | |
| Dense | 20.16 | 51.86 | 30.30 | 35.47 | 63.01 | 52.60 | 88.80 | 64.94 | 70.74 | 93.60 |
| + TF–IDF ($\alpha$) | 22.70 | 57.34 | 32.40 | 38.30 | 66.54 | 62.90 | 94.00 | 73.97 | 78.88 | 97.40 |
| + BM25 ($\alpha$) | 22.11 | 57.34 | 32.54 | 38.43 | 65.95 | 66.70 | 95.40 | 76.74 | 81.30 | 97.90 |
| + TF–IDF (RRF) | 20.55 | 55.19 | 29.80 | 35.75 | 67.12 | 60.10 | 93.80 | 71.95 | 77.28 | 97.50 |
| + BM25 (RRF) | 22.90 | 54.99 | 31.36 | 36.89 | 66.14 | 66.40 | 94.40 | 76.09 | 80.56 | 98.00 |
| AITeamVN/Vietnamese_Embedding_v2 | | | | | | | | | | |
| Dense | 19.96 | 50.29 | 29.11 | 34.17 | 59.88 | 61.40 | 92.20 | 72.04 | 76.96 | 95.60 |
| + TF–IDF ($\alpha$) | 19.18 | 56.36 | 29.42 | 35.77 | 65.36 | 68.90 | 96.00 | 78.93 | 83.14 | 97.70 |
| + BM25 ($\alpha$) | 19.57 | 56.36 | 29.95 | 36.20 | 65.17 | **72.50** | **96.90** | **81.42** | **85.23** | **98.80** |
| + TF–IDF (RRF) | 19.37 | 53.82 | 28.84 | 34.73 | 65.36 | 63.80 | 95.80 | 75.41 | 80.43 | 97.80 |
| + BM25 (RRF) | 21.14 | 53.42 | 29.94 | 35.47 | 64.58 | 68.60 | 96.20 | 78.72 | 83.03 | 98.30 |
| bkai-foundation-models/vietnamese-bi-encoder | | | | | | | | | | |
| Dense | 18.79 | 48.53 | 27.01 | 32.07 | 58.51 | 46.80 | 77.80 | 56.58 | 61.67 | 85.50 |
| + TF–IDF ($\alpha$) | 19.18 | 52.25 | 28.58 | 34.19 | 64.97 | 59.00 | 91.30 | 69.89 | 75.08 | 95.20 |
| + BM25 ($\alpha$) | 20.74 | 52.84 | 29.95 | 35.38 | 64.19 | 62.10 | 91.90 | 72.34 | 77.09 | 96.00 |
| + TF–IDF (RRF) | 20.16 | 52.05 | 28.53 | 34.05 | 64.77 | 55.30 | 90.00 | 66.77 | 72.38 | 95.90 |
| + BM25 (RRF) | 20.94 | 52.64 | 29.52 | 34.97 | 64.97 | 59.50 | 90.70 | 70.00 | 75.01 | 96.10 |
| dangvantuan/vietnamese-document-embedding | | | | | | | | | | |
| Dense | 20.55 | 53.42 | 30.76 | 36.19 | 63.21 | 50.80 | 86.70 | 63.01 | 68.75 | 92.20 |
| + TF–IDF ($\alpha$) | 20.94 | 56.75 | 31.15 | 37.21 | 66.14 | 62.50 | 93.40 | 73.35 | 78.24 | 96.90 |
| + BM25 ($\alpha$) | 21.92 | 56.36 | 31.91 | 37.74 | 65.56 | 66.70 | 94.90 | 76.34 | 80.85 | 97.20 |
| + TF–IDF (RRF) | 20.16 | 54.99 | 28.89 | 34.97 | 64.97 | 58.30 | 92.80 | 70.36 | 75.83 | 96.90 |
| + BM25 (RRF) | 19.96 | 54.79 | 29.33 | 35.31 | 64.77 | 64.00 | 94.20 | 74.50 | 79.29 | 97.50 |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | | | | | | | | | | |
| Dense | 14.48 | 40.70 | 20.90 | 25.49 | 52.05 | 30.80 | 63.00 | 40.43 | 45.81 | 71.90 |
| + TF–IDF ($\alpha$) | 18.79 | 50.68 | 27.07 | 32.59 | 61.25 | 52.10 | 84.80 | 62.76 | 68.07 | 90.60 |
| + BM25 ($\alpha$) | 18.59 | 52.84 | 27.68 | 33.57 | 61.45 | 55.10 | 85.70 | 65.18 | 70.13 | 92.70 |
| + TF–IDF (RRF) | 17.81 | 51.47 | 25.98 | 31.88 | 61.84 | 44.20 | 80.90 | 56.02 | 62.01 | 90.60 |
| + BM25 (RRF) | 18.59 | 52.84 | 27.06 | 33.03 | 61.64 | 47.00 | 82.50 | 58.78 | 64.52 | 92.50 |
| intfloat/multilingual-e5-large | | | | | | | | | | |
| Dense | 23.87 | 53.82 | 32.70 | 37.72 | 65.56 | 58.50 | 91.90 | 69.85 | 75.21 | 95.60 |
| + TF–IDF ($\alpha$) | 23.29 | **61.06** | 33.62 | 40.08 | 68.88 | 65.40 | 94.50 | 75.96 | 80.52 | 97.60 |
| + BM25 ($\alpha$) | **25.24** | 60.47 | **35.24** | **41.19** | **70.06** | 69.60 | 95.50 | 79.19 | 83.21 | 97.70 |
| + TF–IDF (RRF) | 23.48 | 59.10 | 32.62 | 38.81 | 68.69 | 61.80 | 93.50 | 73.21 | 78.18 | 96.90 |
| + BM25 (RRF) | 21.92 | 57.73 | 32.41 | 38.42 | 69.08 | 66.70 | 94.90 | 77.06 | 81.45 | 97.10 |
| jinaai/jina-embeddings-v3 | | | | | | | | | | |
| Dense | 21.72 | 52.05 | 31.01 | 36.02 | 62.43 | 49.50 | 86.10 | 61.08 | 67.09 | 91.30 |
| + TF–IDF ($\alpha$) | 22.11 | 58.32 | 32.34 | 38.49 | 68.30 | 59.60 | 92.30 | 70.88 | 76.10 | 96.50 |
| + BM25 ($\alpha$) | 23.48 | 59.30 | 34.13 | 40.12 | 67.91 | 64.10 | 93.70 | 74.08 | 78.83 | 97.00 |
| + TF–IDF (RRF) | 21.33 | 56.36 | 30.61 | 36.66 | 69.28 | 57.00 | 90.50 | 68.39 | 73.76 | 96.50 |
| + BM25 (RRF) | 20.74 | 56.75 | 30.81 | 36.93 | 69.28 | 60.40 | 93.40 | 71.73 | 77.00 | 96.90 |
| BAAI/bge-m3 | | | | | | | | | | |
| Dense | 24.66 | 55.77 | 34.22 | 39.37 | 64.77 | 59.20 | 91.10 | 70.40 | 75.45 | 95.10 |
| + TF–IDF ($\alpha$) | 23.68 | 57.93 | 33.78 | 39.52 | 67.51 | 65.90 | 95.30 | 76.55 | 81.15 | 97.40 |
| + BM25 ($\alpha$) | 22.90 | 58.71 | 33.68 | 39.63 | 67.51 | 71.10 | 95.90 | 79.94 | 83.86 | 98.20 |
| + TF–IDF (RRF) | 20.55 | 53.82 | 29.72 | 35.40 | 67.32 | 61.80 | 94.60 | 73.72 | 78.85 | 97.70 |
| + BM25 (RRF) | 21.92 | 55.38 | 31.13 | 36.83 | 67.12 | 67.60 | 95.60 | 77.43 | 81.86 | 97.60 |
| Snowflake/snowflake-arctic-embed-l-v2.0 | | | | | | | | | | |
| Dense | 20.74 | 54.79 | 31.05 | 36.73 | 64.58 | 60.00 | 91.50 | 70.41 | 75.50 | 95.30 |
| + TF–IDF ($\alpha$) | 22.50 | 58.90 | 32.47 | 38.69 | 68.10 | 63.40 | 93.90 | 74.50 | 79.26 | 97.20 |
| + BM25 ($\alpha$) | 23.48 | 58.51 | 33.33 | 39.28 | 68.49 | 68.20 | 95.70 | 78.02 | 82.35 | 97.70 |

Continued on next page

Table 12 – continued from previous page

| Method | EduCoQA | | | | | ViRHE4QA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| + TF–IDF (RRF) | 19.96 | 54.99 | 29.70 | 35.66 | 67.91 | 60.80 | 92.90 | 72.12 | 77.20 | 96.40 |
| + BM25 (RRF) | 20.35 | 55.77 | 30.49 | 36.48 | 66.54 | 65.80 | 94.90 | 76.31 | 80.87 | 97.50 |
| 🤗 Alibaba-NLP/gte-multilingual-base | | | | | | | | | | |
| Dense | 18.00 | 52.45 | 28.72 | 34.42 | 62.82 | 54.40 | 87.80 | 65.52 | 70.91 | 93.10 |
| + TF–IDF ($\alpha$) | 21.53 | 59.10 | 31.79 | 38.23 | 68.49 | 61.10 | 93.80 | 72.66 | 77.81 | 96.20 |
| + BM25 ($\alpha$) | 21.53 | 59.69 | 32.41 | 38.88 | 69.28 | 65.50 | 94.50 | 75.77 | 80.33 | 97.10 |
| + TF–IDF (RRF) | 18.98 | 54.99 | 29.02 | 35.17 | 69.47 | 59.10 | 91.80 | 70.59 | 75.77 | 95.70 |
| + BM25 (RRF) | 19.37 | 57.34 | 29.98 | 36.43 | 68.69 | 62.70 | 93.50 | 73.71 | 78.54 | 97.00 |
| 🤗 BAAI/bge-multilingual-gemma2 | | | | | | | | | | |
| Dense | 12.13 | 40.90 | 20.03 | 24.95 | 51.08 | 50.80 | 82.70 | 61.73 | 66.84 | 89.10 |
| + TF–IDF ($\alpha$) | 19.18 | 46.18 | 26.40 | 31.06 | 57.53 | 60.90 | 93.10 | 72.27 | 77.35 | 96.70 |
| + BM25 ($\alpha$) | 19.18 | 47.16 | 27.13 | 31.88 | 57.93 | 66.20 | 93.70 | 75.89 | 80.23 | 97.60 |
| + TF–IDF (RRF) | 17.81 | 46.58 | 25.81 | 30.71 | 56.75 | 57.40 | 91.80 | 69.41 | 74.86 | 96.90 |
| + BM25 (RRF) | 18.40 | 46.38 | 26.21 | 30.97 | 59.10 | 62.90 | 93.60 | 73.27 | 78.20 | 97.10 |
| 🤗 google/EmbeddingGemma-300m | | | | | | | | | | |
| Dense | 21.14 | 53.82 | 30.70 | 36.19 | 62.23 | 55.20 | 88.50 | 66.37 | 71.73 | 92.80 |
| + TF–IDF ($\alpha$) | 22.31 | 58.12 | 32.23 | 38.35 | 66.14 | 62.60 | 92.80 | 73.34 | 78.11 | 96.70 |
| + BM25 ($\alpha$) | 22.50 | 57.73 | 32.81 | 38.73 | 65.75 | 65.80 | 94.40 | 76.20 | 80.67 | 96.90 |
| + TF–IDF (RRF) | 20.94 | 56.75 | 31.12 | 37.17 | 66.93 | 59.40 | 91.80 | 70.81 | 75.93 | 96.30 |
| + BM25 (RRF) | 20.16 | 57.53 | 30.89 | 37.20 | 65.95 | 63.30 | 94.20 | 74.34 | 79.20 | 96.70 |
| 🤗 Alibaba-NLP/gte-Qwen2-1.5B-instruct | | | | | | | | | | |
| Dense | 6.26 | 21.92 | 9.98 | 12.75 | 29.55 | 31.90 | 68.60 | 42.71 | 48.87 | 77.30 |
| + TF–IDF ($\alpha$) | 11.94 | 38.75 | 19.55 | 24.08 | 46.77 | 55.90 | 90.80 | 67.73 | 73.34 | 94.20 |
| + BM25 ($\alpha$) | 13.50 | 38.16 | 20.55 | 24.71 | 47.55 | 64.40 | 92.20 | 74.30 | 78.69 | 95.60 |
| + TF–IDF (RRF) | 9.39 | 30.72 | 15.07 | 18.72 | 44.03 | 46.70 | 84.10 | 58.92 | 64.99 | 91.70 |
| + BM25 (RRF) | 10.18 | 28.96 | 15.67 | 18.80 | 42.07 | 51.90 | 86.20 | 62.79 | 68.41 | 92.40 |
| 🤗 Qwen/Qwen3-Embedding-0.6B | | | | | | | | | | |
| Dense | 21.53 | 54.99 | 31.50 | 37.10 | 65.95 | 55.40 | 87.80 | 66.28 | 71.48 | 93.20 |
| + TF–IDF ($\alpha$) | 21.72 | 59.69 | 32.47 | 38.92 | 69.67 | 60.80 | 93.00 | 72.12 | 77.22 | 96.90 |
| + BM25 ($\alpha$) | 22.70 | 60.08 | 33.58 | 39.88 | 69.28 | 65.90 | 93.90 | 75.93 | 80.33 | 96.90 |
| + TF–IDF (RRF) | 19.77 | 58.32 | 30.32 | 36.92 | 68.30 | 58.70 | 92.00 | 70.37 | 75.64 | 96.70 |
| + BM25 (RRF) | 20.55 | 59.30 | 31.33 | 37.92 | 67.71 | 63.90 | 93.40 | 74.44 | 79.08 | 97.20 |

Table 13: Retrieval results on **Customer Support Domain** (CSConDa).

| Method | CSConDa | | | | |
|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| TF–IDF | 15.70 | 38.50 | 22.49 | 26.30 | 47.20 |
| BM25 | 17.40 | 36.80 | 22.99 | 26.27 | 45.90 |
| ColBERT | 11.20 | 28.30 | 15.79 | 18.72 | 35.00 |
| SPLADE | 9.10 | 22.20 | 12.65 | 14.89 | 27.70 |
| text-embedding-3-large | | | | | |
| Dense | 33.70 | 56.80 | 41.06 | 44.84 | 63.80 |
| + TF–IDF ($\alpha$) | 34.90 | 60.40 | 42.45 | 46.73 | 66.50 |
| + BM25 ($\alpha$) | **36.40** | 60.20 | **43.60** | 47.55 | 66.20 |
| + TF–IDF (RRF) | 28.80 | 55.00 | 36.45 | 40.85 | 63.80 |
| + BM25 (RRF) | 29.60 | 54.40 | 37.05 | 41.18 | 64.00 |
| 🤗 AITeamVN/Vietnamese_Embedding_v2 | | | | | |
| Dense | 31.40 | 54.00 | 38.40 | 42.14 | 61.40 |
| + TF–IDF ($\alpha$) | 32.70 | 57.50 | 40.51 | 44.59 | 65.30 |
| + BM25 ($\alpha$) | 33.70 | 57.90 | 41.22 | 45.20 | 64.30 |
| + TF–IDF (RRF) | 28.10 | 54.60 | 35.84 | 40.29 | 63.90 |
| + BM25 (RRF) | 28.80 | 54.70 | 36.70 | 40.99 | 62.40 |
| 🤗 bkai-foundation-models/vietnamese-bi-encoder | | | | | |
| Dense | 15.70 | 34.90 | 21.09 | 24.34 | 41.70 |
| + TF–IDF ($\alpha$) | 22.20 | 46.00 | 28.95 | 32.96 | 52.40 |

Continued on next page

Table 13 – continued from previous page

| Method | CSConDa | | | | |
|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| + BM25 ($\alpha$) | 22.70 | 45.70 | 28.93 | 32.87 | 52.80 |
| + TF–IDF (RRF) | 19.10 | 44.20 | 26.34 | 30.56 | 55.00 |
| + BM25 (RRF) | 20.00 | 44.90 | 26.98 | 31.19 | 54.30 |
| 🤗 dangvantuan/vietnamese-document-embedding | | | | | |
| Dense | 28.40 | 53.00 | 36.12 | 40.18 | 59.90 |
| + TF–IDF ($\alpha$) | 31.10 | 57.90 | 39.46 | 43.87 | 65.50 |
| + BM25 ($\alpha$) | 32.40 | 57.80 | 40.05 | 44.28 | 64.60 |
| + TF–IDF (RRF) | 26.00 | 51.80 | 33.88 | 38.16 | 64.00 |
| + BM25 (RRF) | 26.70 | 52.60 | 34.65 | 38.92 | 63.70 |
| 🤗 sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | | | | | |
| Dense | 11.80 | 30.00 | 16.71 | 19.82 | 39.10 |
| + TF–IDF ($\alpha$) | 19.60 | 45.30 | 27.05 | 31.38 | 51.40 |
| + BM25 ($\alpha$) | 18.90 | 45.40 | 26.18 | 30.71 | 51.60 |
| + TF–IDF (RRF) | 17.70 | 43.80 | 25.03 | 29.47 | 54.30 |
| + BM25 (RRF) | 17.50 | 43.80 | 24.63 | 29.15 | 53.30 |
| 🤗 intfloat/multilingual-e5-large | | | | | |
| Dense | 27.20 | 48.10 | 33.78 | 37.21 | 55.90 |
| + TF–IDF ($\alpha$) | 31.90 | 53.60 | 38.76 | 42.32 | 61.70 |
| + BM25 ($\alpha$) | 32.60 | 54.90 | 39.46 | 43.16 | 60.70 |
| + TF–IDF (RRF) | 28.70 | 53.40 | 36.11 | 40.22 | 61.40 |
| + BM25 (RRF) | 29.00 | 53.50 | 36.30 | 40.39 | 61.30 |
| 🤗 jinaai/jina-embeddings-v3 | | | | | |
| Dense | 32.10 | 57.40 | 40.03 | 44.19 | 64.30 |
| + TF–IDF ($\alpha$) | 34.90 | **61.20** | 42.70 | 47.12 | 66.90 |
| + BM25 ($\alpha$) | 35.40 | **61.20** | 43.42 | **47.68** | **67.60** |
| + TF–IDF (RRF) | 29.30 | 57.30 | 37.96 | 42.58 | 66.10 |
| + BM25 (RRF) | 31.40 | 56.60 | 39.01 | 43.20 | 65.90 |
| 🤗 BAAI/bge-m3 | | | | | |
| Dense | 30.80 | 53.90 | 37.98 | 41.80 | 61.00 |
| + TF–IDF ($\alpha$) | 33.10 | 57.00 | 40.67 | 44.59 | 63.80 |
| + BM25 ($\alpha$) | 33.90 | 56.90 | 40.97 | 44.78 | 63.90 |
| + TF–IDF (RRF) | 28.40 | 54.20 | 35.82 | 40.17 | 63.90 |
| + BM25 (RRF) | 28.60 | 53.70 | 36.24 | 40.40 | 62.80 |
| 🤗 Snowflake/snowflake-arctic-embed-l-v2.0 | | | | | |
| Dense | 32.80 | 56.70 | 40.69 | 44.56 | 63.20 |
| + TF–IDF ($\alpha$) | 34.30 | 58.80 | 42.15 | 46.15 | 66.20 |
| + BM25 ($\alpha$) | 35.60 | 59.80 | 43.56 | 47.46 | 66.30 |
| + TF–IDF (RRF) | 30.10 | 55.30 | 37.85 | 42.03 | 65.00 |
| + BM25 (RRF) | 29.80 | 56.40 | 38.16 | 42.54 | 65.20 |
| 🤗 Alibaba-NLP/gte-multilingual-base | | | | | |
| Dense | 28.10 | 51.60 | 35.23 | 39.13 | 57.70 |
| + TF–IDF ($\alpha$) | 29.30 | 55.50 | 37.45 | 41.78 | 63.50 |
| + BM25 ($\alpha$) | 30.80 | 56.80 | 38.82 | 43.12 | 63.20 |
| + TF–IDF (RRF) | 26.70 | 52.10 | 34.42 | 38.64 | 62.10 |
| + BM25 (RRF) | 27.20 | 52.00 | 34.88 | 38.98 | 62.20 |
| 🤗 BAAI/bge-multilingual-gemma2 | | | | | |
| Dense | 14.30 | 30.50 | 18.65 | 21.43 | 37.00 |
| + TF–IDF ($\alpha$) | 23.50 | 43.00 | 29.72 | 32.91 | 49.60 |
| + BM25 ($\alpha$) | 23.30 | 43.30 | 29.50 | 32.81 | 49.90 |
| + TF–IDF (RRF) | 19.10 | 43.30 | 25.92 | 30.01 | 52.20 |
| + BM25 (RRF) | 20.30 | 43.10 | 27.02 | 30.84 | 51.80 |
| 🤗 google/EmbeddingGemma-300m | | | | | |
| Dense | 29.90 | 54.70 | 37.34 | 41.49 | 60.80 |
| + TF–IDF ($\alpha$) | 32.50 | 57.80 | 40.50 | 44.65 | 65.10 |
| + BM25 ($\alpha$) | 33.60 | 58.80 | 41.28 | 45.45 | 64.90 |
| + TF–IDF (RRF) | 28.30 | 54.60 | 36.23 | 40.61 | 63.20 |
| + BM25 (RRF) | 28.70 | 54.50 | 36.35 | 40.67 | 63.30 |
| 🤗 Alibaba-NLP/gte-Qwen2-1.5B-instruct | | | | | |
| Dense | 6.70 | 16.50 | 9.33 | 11.01 | 21.20 |

Continued on next page

Table 13 – continued from previous page

| Method | CSConDa | | | | |
|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| + TF–IDF ($\alpha$) | 17.40 | 35.80 | 22.82 | 25.90 | 42.20 |
| + BM25 ($\alpha$) | 16.40 | 34.70 | 21.68 | 24.76 | 41.00 |
| + TF–IDF (RRF) | 11.90 | 32.10 | 17.29 | 20.76 | 44.50 |
| + BM25 (RRF) | 12.00 | 32.00 | 17.29 | 20.72 | 43.80 |
| Qwen/Qwen3-Embedding-0.6B | | | | | |
| Dense | 24.80 | 49.40 | 32.27 | 36.34 | 56.60 |
| + TF–IDF ($\alpha$) | 29.10 | 56.40 | 37.31 | 41.85 | 63.10 |
| + BM25 ($\alpha$) | 30.10 | 55.70 | 38.11 | 42.33 | 62.40 |
| + TF–IDF (RRF) | 24.60 | 53.20 | 33.14 | 37.91 | 62.20 |
| + BM25 (RRF) | 26.70 | 53.10 | 34.41 | 38.85 | 62.30 |

Table 14: Retrieval results on **Legal Domain** (ALQAC and ZaloLegalQA).

| Method | ALQAC | | | | | ZaloLegalQA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| TF–IDF | 82.83 | 96.23 | 88.34 | 90.31 | 98.11 | 64.70 | 92.47 | 75.07 | 79.28 | 96.45 |
| BM25 | 89.25 | 97.92 | 92.20 | 93.59 | 99.25 | 71.40 | 92.18 | 79.39 | 82.49 | 94.42 |
| ColBERT | 69.43 | 89.62 | 76.42 | 79.63 | 93.40 | 54.40 | 75.98 | 61.78 | 65.07 | 80.73 |
| SPLADE | 68.68 | 91.32 | 76.33 | 79.97 | 95.66 | 40.30 | 65.53 | 48.42 | 52.35 | 72.58 |
| text-embedding-3-large | | | | | | | | | | |
| Dense | 84.53 | 98.68 | 90.13 | 92.26 | 99.81 | 80.50 | 96.83 | 87.10 | 89.39 | 98.53 |
| + TF–IDF ($\alpha$) | 88.87 | 99.25 | 93.22 | 94.74 | 99.81 | 82.20 | 98.07 | 88.52 | 90.80 | 99.27 |
| + BM25 ($\alpha$) | 93.02 | 99.43 | 95.79 | 96.72 | **100.00** | 84.00 | 97.93 | 89.83 | 91.77 | 98.92 |
| + TF–IDF (RRF) | 87.36 | 98.30 | 91.83 | 93.45 | 99.06 | 77.10 | 97.72 | 85.08 | 88.15 | 98.90 |
| + BM25 (RRF) | 90.57 | 98.87 | 94.06 | 95.28 | 99.81 | 78.10 | 96.28 | 85.50 | 88.13 | 98.12 |
| AITeamVN/Vietnamese_Embedding_v2 | | | | | | | | | | |
| Dense | 90.38 | 99.06 | 93.96 | 95.24 | 99.81 | 86.20 | 98.32 | 91.04 | 92.74 | 98.97 |
| + TF–IDF ($\alpha$) | 92.64 | 99.62 | 95.26 | 96.33 | 99.81 | 85.40 | 98.42 | 90.91 | 92.71 | 99.10 |
| + BM25 ($\alpha$) | 93.77 | 99.25 | 95.95 | 96.78 | **100.00** | **87.40** | 98.17 | **91.67** | **93.18** | 98.67 |
| + TF–IDF (RRF) | 90.19 | 98.49 | 93.15 | 94.44 | 99.81 | 79.10 | 97.65 | 86.65 | 89.35 | 98.60 |
| + BM25 (RRF) | 92.08 | 98.68 | 94.78 | 95.75 | 99.81 | 81.00 | 96.27 | 87.12 | 89.31 | 97.75 |
| bkai-foundation-models/vietnamese-bi-encoder | | | | | | | | | | |
| Dense | 80.75 | 95.66 | 86.21 | 88.52 | 98.11 | 71.00 | 92.72 | 79.20 | 82.46 | 94.72 |
| + TF–IDF ($\alpha$) | 88.68 | 98.49 | 92.60 | 94.07 | 99.25 | 77.40 | 96.72 | 84.88 | 87.76 | 98.20 |
| + BM25 ($\alpha$) | 89.62 | 98.68 | 93.33 | 94.67 | 99.62 | 80.30 | 96.67 | 86.67 | 89.08 | 97.97 |
| + TF–IDF (RRF) | 84.53 | 97.74 | 89.43 | 91.48 | 99.25 | 72.90 | 95.52 | 81.48 | 84.88 | 97.85 |
| + BM25 (RRF) | 88.30 | 98.30 | 92.13 | 93.65 | 99.62 | 75.30 | 95.62 | 82.96 | 85.99 | 97.20 |
| dangvantuan/vietnamese-document-embedding | | | | | | | | | | |
| Dense | 85.85 | 98.49 | 90.51 | 92.48 | 99.25 | 77.70 | 96.27 | 85.08 | 87.81 | 97.95 |
| + TF–IDF ($\alpha$) | 89.06 | 98.87 | 92.96 | 94.44 | 99.62 | 79.80 | 98.35 | 86.80 | 89.60 | 99.00 |
| + BM25 ($\alpha$) | 92.64 | 99.25 | 95.21 | 96.21 | 99.62 | 82.40 | 97.42 | 88.33 | 90.54 | 98.60 |
| + TF–IDF (RRF) | 87.92 | 97.92 | 91.80 | 93.32 | 99.43 | 76.20 | 97.10 | 84.15 | 87.31 | 98.50 |
| + BM25 (RRF) | 89.43 | 98.68 | 93.28 | 94.63 | 99.43 | 77.60 | 96.30 | 84.92 | 87.68 | 97.80 |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | | | | | | | | | | |
| Dense | 65.85 | 92.45 | 74.37 | 78.73 | 95.09 | 51.10 | 80.23 | 60.75 | 65.30 | 85.87 |
| + TF–IDF ($\alpha$) | 83.96 | 97.36 | 89.27 | 91.29 | 99.06 | 69.70 | 93.92 | 78.50 | 82.18 | 96.60 |
| + BM25 ($\alpha$) | 88.30 | 97.36 | 91.80 | 93.18 | 99.25 | 70.60 | 94.32 | 79.43 | 82.98 | 95.95 |
| + TF–IDF (RRF) | 77.55 | 96.79 | 84.64 | 87.62 | 98.68 | 64.10 | 91.60 | 73.74 | 78.00 | 96.25 |
| + BM25 (RRF) | 80.19 | 97.55 | 86.93 | 89.55 | 98.68 | 64.60 | 91.62 | 74.46 | 78.56 | 96.70 |
| intfloat/multilingual-e5-large | | | | | | | | | | |
| Dense | 89.06 | 99.25 | 92.91 | 94.48 | 99.81 | 83.80 | 97.98 | 89.69 | 91.64 | 98.88 |
| + TF–IDF ($\alpha$) | 89.81 | 98.87 | 93.70 | 95.01 | 99.81 | 83.80 | 98.48 | 89.56 | 91.71 | 98.77 |
| + BM25 ($\alpha$) | 92.64 | 99.43 | 95.47 | 96.47 | **100.00** | 84.80 | 98.08 | 90.15 | 92.05 | 98.53 |
| + TF–IDF (RRF) | 86.98 | 98.87 | 91.84 | 93.60 | 99.81 | 76.40 | 97.58 | 84.82 | 87.92 | 98.42 |
| + BM25 (RRF) | 91.32 | 99.25 | 94.73 | 95.87 | 99.81 | 78.20 | 95.33 | 84.86 | 87.41 | 97.57 |
| jinaai/jina-embeddings-v3 | | | | | | | | | | |

Continued on next page

Table 14 – continued from previous page

| Method | ALQAC | | | | | ZaloLegalQA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| Dense | 84.34 | 98.49 | 89.64 | 91.83 | 99.62 | 83.10 | 98.85 | 89.39 | 91.67 | **99.40** |
| + TF–IDF ($\alpha$) | 87.92 | 99.06 | 92.65 | 94.27 | **100.00** | 81.40 | **98.90** | 88.36 | 90.94 | 99.25 |
| + BM25 ($\alpha$) | 90.38 | 99.62 | 94.12 | 95.49 | 99.81 | 85.00 | 98.55 | 90.43 | 92.40 | 99.15 |
| + TF–IDF (RRF) | 85.28 | 99.06 | 90.99 | 93.01 | **100.00** | 76.40 | 97.75 | 84.77 | 87.95 | 98.90 |
| + BM25 (RRF) | 89.62 | 99.25 | 93.95 | 95.29 | 99.81 | 78.80 | 96.17 | 85.27 | 87.92 | 97.65 |
| 🤗 BAAI/bge-m3 | | | | | | | | | | |
| Dense | 90.38 | 99.43 | 94.14 | 95.47 | **100.00** | 82.30 | 97.83 | 88.52 | 90.75 | 98.77 |
| + TF–IDF ($\alpha$) | 92.08 | 99.43 | 95.02 | 96.12 | 99.81 | 80.50 | 98.12 | 87.69 | 90.26 | 99.05 |
| + BM25 ($\alpha$) | **94.72** | 99.62 | **96.66** | **97.40** | 99.81 | 82.10 | 98.02 | 88.62 | 90.93 | 98.62 |
| + TF–IDF (RRF) | 88.49 | 98.49 | 92.37 | 93.88 | 99.62 | 76.00 | 97.22 | 84.43 | 87.58 | 98.70 |
| + BM25 (RRF) | 91.13 | 99.06 | 94.40 | 95.56 | 99.62 | 79.20 | 96.17 | 86.05 | 88.52 | 97.67 |
| 🤗 Snowflake/snowflake-arctic-embed-l-v2.0 | | | | | | | | | | |
| Dense | 88.30 | 99.43 | 92.68 | 94.36 | 99.81 | 83.20 | 98.03 | 89.30 | 91.42 | 98.72 |
| + TF–IDF ($\alpha$) | 89.81 | 99.25 | 93.51 | 94.94 | 99.62 | 82.00 | 98.12 | 88.28 | 90.68 | 98.47 |
| + BM25 ($\alpha$) | 92.83 | 99.62 | 95.62 | 96.62 | 99.81 | 84.90 | 98.13 | 90.07 | 91.99 | 98.42 |
| + TF–IDF (RRF) | 87.17 | 99.06 | 91.76 | 93.57 | 99.62 | 76.80 | 97.02 | 84.59 | 87.63 | 98.42 |
| + BM25 (RRF) | 91.70 | 99.25 | 94.89 | 95.99 | 99.62 | 77.70 | 95.33 | 84.63 | 87.23 | 97.32 |
| 🤗 Alibaba-NLP/gte-multilingual-base | | | | | | | | | | |
| Dense | 88.30 | 98.30 | 92.04 | 93.59 | 99.62 | 79.00 | 96.22 | 85.56 | 88.13 | 98.32 |
| + TF–IDF ($\alpha$) | 90.00 | 98.87 | 93.46 | 94.81 | 99.81 | 79.10 | 97.82 | 86.12 | 88.94 | 98.40 |
| + BM25 ($\alpha$) | 91.89 | 99.06 | 94.90 | 95.94 | 99.81 | 80.90 | 97.17 | 87.09 | 89.52 | 97.97 |
| + TF–IDF (RRF) | 88.11 | 98.68 | 92.06 | 93.70 | **100.00** | 74.20 | 97.05 | 82.97 | 86.40 | 98.25 |
| + BM25 (RRF) | 90.38 | 99.25 | 93.84 | 95.18 | 99.81 | 76.60 | 95.27 | 83.68 | 86.48 | 97.07 |
| 🤗 BAAI/bge-multilingual-gemma2 | | | | | | | | | | |
| Dense | 85.28 | 99.06 | 90.37 | 92.51 | 99.43 | 75.60 | 95.68 | 83.21 | 86.18 | 97.15 |
| + TF–IDF ($\alpha$) | 88.87 | 99.06 | 93.20 | 94.67 | 99.43 | 79.10 | 97.75 | 86.53 | 89.26 | 98.90 |
| + BM25 ($\alpha$) | 92.26 | 99.62 | 95.22 | 96.32 | 99.81 | 82.60 | 97.72 | 88.60 | 90.78 | 98.60 |
| + TF–IDF (RRF) | 88.11 | 98.11 | 92.31 | 93.76 | 99.06 | 75.60 | 96.50 | 83.77 | 86.87 | 98.60 |
| + BM25 (RRF) | 92.64 | 98.87 | 95.01 | 95.97 | 99.43 | 76.50 | 96.10 | 84.47 | 87.28 | 97.85 |
| 🤗 google/EmbeddingGemma-300m | | | | | | | | | | |
| Dense | 87.74 | 99.06 | 92.07 | 93.80 | 99.43 | 82.50 | 96.67 | 88.11 | 90.12 | 97.87 |
| + TF–IDF ($\alpha$) | 89.25 | 99.43 | 93.20 | 94.74 | 99.81 | 80.60 | 97.87 | 87.42 | 89.94 | 98.57 |
| + BM25 ($\alpha$) | 91.32 | **100.00** | 94.61 | 95.93 | **100.00** | 83.00 | 97.37 | 88.65 | 90.71 | 98.22 |
| + TF–IDF (RRF) | 87.17 | 99.06 | 91.70 | 93.52 | 99.81 | 76.00 | 96.72 | 84.19 | 87.23 | 98.27 |
| + BM25 (RRF) | 90.38 | 99.62 | 94.00 | 95.40 | 99.81 | 77.90 | 94.97 | 84.39 | 86.94 | 96.97 |
| 🤗 Alibaba-NLP/gte-Qwen2-1.5B-instruct | | | | | | | | | | |
| Dense | 65.47 | 89.62 | 73.38 | 77.29 | 94.34 | 84.10 | 98.23 | 89.76 | 91.77 | 99.05 |
| + TF–IDF ($\alpha$) | 82.26 | 96.79 | 88.03 | 90.21 | 98.11 | 82.50 | 98.60 | 88.79 | 91.19 | 99.05 |
| + BM25 ($\alpha$) | 87.92 | 97.55 | 91.67 | 93.13 | 98.11 | 86.00 | 98.47 | 90.99 | 92.76 | 98.95 |
| + TF–IDF (RRF) | 76.23 | 95.66 | 83.16 | 86.22 | 98.11 | 75.90 | 97.25 | 84.42 | 87.58 | 98.70 |
| + BM25 (RRF) | 79.06 | 96.60 | 85.19 | 87.97 | 98.49 | 78.20 | 95.67 | 85.03 | 87.62 | 97.90 |
| 🤗 Qwen/Qwen3-Embedding-0.6B | | | | | | | | | | |
| Dense | 85.66 | 99.06 | 90.44 | 92.54 | 99.43 | 80.80 | 97.68 | 87.48 | 89.88 | 98.40 |
| + TF–IDF ($\alpha$) | 89.62 | 98.68 | 93.05 | 94.44 | 99.81 | 80.10 | 98.20 | 87.15 | 89.82 | 98.65 |
| + BM25 ($\alpha$) | 91.51 | 99.81 | 94.60 | 95.88 | 99.81 | 83.20 | 97.63 | 88.90 | 90.96 | 98.45 |
| + TF–IDF (RRF) | 86.42 | 98.68 | 91.25 | 93.09 | 99.62 | 77.10 | 97.25 | 84.84 | 87.86 | 98.35 |
| + BM25 (RRF) | 89.81 | 99.43 | 93.71 | 95.14 | 99.62 | 77.60 | 95.27 | 84.47 | 87.09 | 97.45 |

Table 15: Retrieval results on **Healthcare Domain** (ViNewsQA and ViMedAQA).

| Method | ViNewsQA | | | | | ViMedAQA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| TF–IDF | 52.20 | 79.10 | 60.93 | 65.31 | 84.70 | 61.50 | 84.80 | 69.46 | 73.18 | 88.20 |
| BM25 | 59.00 | 80.20 | 66.13 | 69.53 | 84.30 | 65.40 | 84.50 | 71.36 | 74.51 | 87.30 |
| ColBERT | 29.70 | 51.00 | 35.77 | 39.37 | 57.50 | 49.60 | 71.40 | 56.42 | 60.01 | 75.50 |
| SPLADE | 26.70 | 50.20 | 33.44 | 37.40 | 58.00 | 47.40 | 68.80 | 54.08 | 57.60 | 73.60 |

Continued on next page

Table 15 – continued from previous page

| Method | ViNewsQA | | | | | ViMedAQA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| text-embedding-3-large | | | | | | | | | | |
| Dense | 49.20 | 76.40 | 58.59 | 62.90 | 81.40 | 80.40 | 95.40 | 85.44 | 87.84 | 97.20 |
| + TF–IDF ($\alpha$) | 62.40 | 84.50 | 69.96 | 73.49 | 89.60 | 80.90 | 96.30 | 86.22 | 88.66 | 97.80 |
| + BM25 ($\alpha$) | 64.70 | 85.90 | 71.83 | 75.23 | 90.80 | 83.00 | 96.10 | 87.37 | 89.48 | 97.90 |
| + TF–IDF (RRF) | 56.20 | 83.30 | 65.25 | 69.63 | 89.40 | 74.70 | 91.70 | 80.63 | 83.33 | 95.20 |
| + BM25 (RRF) | 61.00 | 85.30 | 68.74 | 72.72 | 90.20 | 77.70 | 92.10 | 82.16 | 84.53 | 95.00 |
| AITeamVN/Vietnamese_Embedding_v2 | | | | | | | | | | |
| Dense | 55.40 | 78.20 | 63.01 | 66.69 | 82.90 | 76.90 | 93.50 | 82.40 | 85.06 | 96.10 |
| + TF–IDF ($\alpha$) | 65.30 | 86.10 | 72.35 | 75.68 | 90.00 | 79.80 | 94.10 | 84.61 | 86.91 | 96.20 |
| + BM25 ($\alpha$) | **68.60** | 87.30 | **74.96** | **77.95** | 90.00 | 81.40 | 93.80 | 85.38 | 87.40 | 95.70 |
| + TF–IDF (RRF) | 62.10 | 85.40 | 69.71 | 73.48 | 89.90 | 75.00 | 91.80 | 80.92 | 83.57 | 94.60 |
| + BM25 (RRF) | 65.10 | 86.30 | 72.00 | 75.44 | 89.90 | 76.60 | 90.80 | 81.43 | 83.71 | 94.40 |
| bkai-foundation-models/vietnamese-bi-encoder | | | | | | | | | | |
| Dense | 45.50 | 67.50 | 52.39 | 56.01 | 73.20 | 70.10 | 87.20 | 75.68 | 78.46 | 90.40 |
| + TF–IDF ($\alpha$) | 58.80 | 80.70 | 66.41 | 69.88 | 85.40 | 75.70 | 90.90 | 81.10 | 83.50 | 93.10 |
| + BM25 ($\alpha$) | 59.90 | 80.40 | 66.83 | 70.11 | 85.60 | 77.10 | 90.80 | 82.00 | 84.15 | 92.80 |
| + TF–IDF (RRF) | 52.70 | 78.80 | 61.72 | 65.86 | 86.40 | 69.90 | 89.30 | 76.88 | 79.92 | 92.60 |
| + BM25 (RRF) | 56.90 | 79.60 | 64.12 | 67.83 | 86.80 | 72.80 | 89.30 | 78.34 | 80.99 | 92.60 |
| dangvantuan/vietnamese-document-embedding | | | | | | | | | | |
| Dense | 54.90 | 76.80 | 61.77 | 65.38 | 80.50 | 75.50 | 90.60 | 80.86 | 83.24 | 93.00 |
| + TF–IDF ($\alpha$) | 61.70 | 84.30 | 69.22 | 72.86 | 88.30 | 78.30 | 93.10 | 83.73 | 86.03 | 94.70 |
| + BM25 ($\alpha$) | 64.70 | 85.20 | 71.51 | 74.82 | 88.20 | 79.80 | 92.60 | 84.37 | 86.39 | 94.50 |
| + TF–IDF (RRF) | 57.90 | 82.30 | 66.03 | 69.97 | 88.20 | 73.00 | 91.10 | 79.39 | 82.23 | 94.10 |
| + BM25 (RRF) | 62.70 | 83.80 | 69.44 | 72.90 | 89.80 | 74.70 | 90.40 | 80.03 | 82.54 | 93.50 |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | | | | | | | | | | |
| Dense | 25.50 | 52.20 | 33.54 | 37.98 | 60.00 | 48.00 | 73.50 | 56.19 | 60.34 | 80.70 |
| + TF–IDF ($\alpha$) | 50.70 | 76.80 | 58.94 | 63.23 | 83.10 | 67.00 | 87.10 | 74.05 | 77.22 | 90.60 |
| + BM25 ($\alpha$) | 49.70 | 77.30 | 58.61 | 63.10 | 82.80 | 69.70 | 87.00 | 75.24 | 78.06 | 90.20 |
| + TF–IDF (RRF) | 41.70 | 73.40 | 51.25 | 56.53 | 84.70 | 61.40 | 84.20 | 69.16 | 72.80 | 90.40 |
| + BM25 (RRF) | 44.40 | 76.10 | 53.95 | 59.23 | 85.60 | 63.90 | 83.80 | 70.27 | 73.51 | 89.90 |
| intfloat/multilingual-e5-large | | | | | | | | | | |
| Dense | 57.40 | 79.70 | 64.91 | 68.48 | 84.40 | 83.30 | 96.40 | 87.94 | 90.01 | 97.80 |
| + TF–IDF ($\alpha$) | 64.30 | 85.90 | 71.35 | 74.85 | 89.70 | 82.80 | 95.90 | 87.46 | 89.51 | 97.50 |
| + BM25 ($\alpha$) | 66.10 | **88.50** | 73.59 | 77.19 | **91.60** | **84.70** | **96.50** | **88.82** | **90.69** | 97.70 |
| + TF–IDF (RRF) | 61.40 | 84.80 | 69.07 | 72.87 | 90.10 | 78.40 | 94.20 | 84.38 | 86.80 | 96.30 |
| + BM25 (RRF) | 65.00 | 87.00 | 72.10 | 75.68 | 91.40 | 81.20 | 94.60 | 85.92 | 88.03 | 96.50 |
| jinaai/jina-embeddings-v3 | | | | | | | | | | |
| Dense | 55.00 | 76.40 | 61.95 | 65.44 | 81.70 | 79.90 | 95.40 | 85.29 | 87.75 | 97.30 |
| + TF–IDF ($\alpha$) | 62.40 | 83.50 | 69.46 | 72.86 | 88.20 | 82.50 | 95.80 | 87.39 | 89.45 | 97.80 |
| + BM25 ($\alpha$) | 64.80 | 84.90 | 71.60 | 74.83 | 90.00 | **84.70** | 96.00 | 88.66 | 90.45 | 97.70 |
| + TF–IDF (RRF) | 58.30 | 83.10 | 66.57 | 70.57 | 88.70 | 77.10 | 94.10 | 83.83 | 86.38 | 95.90 |
| + BM25 (RRF) | 63.80 | 85.00 | 70.66 | 74.11 | 90.40 | 79.60 | 94.30 | 85.04 | 87.30 | 96.40 |
| BAAI/bge-m3 | | | | | | | | | | |
| Dense | 57.60 | 79.00 | 64.72 | 68.17 | 83.40 | 81.20 | 94.10 | 85.60 | 87.67 | 96.70 |
| + TF–IDF ($\alpha$) | 63.90 | 86.40 | 71.41 | 75.04 | 89.50 | 81.30 | 94.40 | 85.95 | 88.01 | 96.60 |
| + BM25 ($\alpha$) | 65.50 | 87.10 | 72.87 | 76.33 | 90.10 | 82.50 | 94.60 | 86.58 | 88.52 | 96.70 |
| + TF–IDF (RRF) | 59.90 | 84.60 | 68.10 | 72.08 | 89.50 | 74.70 | 91.40 | 80.76 | 83.36 | 94.80 |
| + BM25 (RRF) | 63.70 | 85.20 | 70.63 | 74.12 | 90.00 | 76.80 | 91.30 | 81.59 | 83.93 | 94.70 |
| Snowflake/snowflake-arctic-embed-l-v2.0 | | | | | | | | | | |
| Dense | 52.70 | 75.00 | 60.12 | 63.71 | 79.70 | 79.40 | 94.30 | 84.56 | 86.93 | 96.00 |
| + TF–IDF ($\alpha$) | 60.20 | 82.30 | 67.62 | 71.18 | 87.00 | 81.20 | 94.70 | 86.03 | 88.15 | 96.50 |
| + BM25 ($\alpha$) | 63.20 | 84.50 | 70.37 | 73.78 | 88.30 | 83.80 | 95.10 | 87.75 | 89.53 | 96.50 |
| + TF–IDF (RRF) | 56.10 | 81.80 | 64.69 | 68.83 | 87.40 | 76.60 | 93.50 | 82.97 | 85.56 | 95.70 |
| + BM25 (RRF) | 61.30 | 83.70 | 68.72 | 72.34 | 89.20 | 80.30 | 93.60 | 84.89 | 87.00 | 95.60 |
| Alibaba-NLP/gte-multilingual-base | | | | | | | | | | |
| Dense | 53.00 | 76.60 | 60.79 | 64.60 | 80.40 | 73.90 | 89.90 | 79.53 | 82.05 | 92.80 |
| + TF–IDF ($\alpha$) | 61.20 | 83.00 | 68.51 | 72.03 | 87.80 | 79.70 | 93.30 | 84.58 | 86.71 | 95.20 |
| + BM25 ($\alpha$) | 63.30 | 85.30 | 70.70 | 74.23 | 89.50 | 80.30 | 93.20 | 84.95 | 86.96 | 95.10 |
| + TF–IDF (RRF) | 58.30 | 82.40 | 65.97 | 69.92 | 88.20 | 74.60 | 91.60 | 80.99 | 83.61 | 94.50 |

Continued on next page

Table 15 – continued from previous page

| Method | ViNewsQA | | | | | ViMedAQA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| + BM25 (RRF) | 62.10 | 84.80 | 69.39 | 73.10 | 89.70 | 76.70 | 92.30 | 82.10 | 84.58 | 94.70 |
| 🤗 BAAI/bge-multilingual-gemma2 | | | | | | | | | | |
| Dense | 44.00 | 70.40 | 52.36 | 56.68 | 75.00 | 58.40 | 80.20 | 65.39 | 68.94 | 86.50 |
| + TF–IDF ($\alpha$) | 58.30 | 82.00 | 66.30 | 70.11 | 86.30 | 73.70 | 92.30 | 80.22 | 83.18 | 94.00 |
| + BM25 ($\alpha$) | 61.80 | 84.20 | 68.99 | 72.65 | 88.70 | 76.20 | 92.60 | 81.99 | 84.59 | 94.80 |
| + TF–IDF (RRF) | 53.50 | 80.00 | 62.29 | 66.57 | 87.20 | 69.50 | 90.60 | 76.46 | 79.87 | 93.90 |
| + BM25 (RRF) | 58.70 | 83.40 | 66.65 | 70.68 | 89.50 | 71.30 | 91.10 | 78.11 | 81.26 | 94.10 |
| 🤗 google/EmbeddingGemma-300m | | | | | | | | | | |
| Dense | 54.30 | 77.90 | 62.21 | 65.99 | 82.70 | 77.80 | 95.30 | 83.62 | 86.45 | 97.30 |
| + TF–IDF ($\alpha$) | 62.40 | 83.80 | 69.50 | 72.95 | 88.10 | 81.20 | 96.10 | 86.50 | 88.85 | **98.20** |
| + BM25 ($\alpha$) | 65.50 | 85.80 | 72.45 | 75.69 | 89.50 | 84.10 | 96.10 | 88.27 | 90.18 | **98.20** |
| + TF–IDF (RRF) | 58.30 | 83.10 | 66.66 | 70.64 | 88.20 | 77.80 | 93.90 | 83.90 | 86.38 | 95.80 |
| + BM25 (RRF) | 62.70 | 85.80 | 70.33 | 74.06 | 91.10 | 80.50 | 93.60 | 85.35 | 87.38 | 96.60 |
| 🤗 Alibaba-NLP/gte-Qwen2-1.5B-instruct | | | | | | | | | | |
| Dense | 54.00 | 75.30 | 61.21 | 64.63 | 80.10 | 59.90 | 81.20 | 66.93 | 70.37 | 84.70 |
| + TF–IDF ($\alpha$) | 61.60 | 82.80 | 68.73 | 72.13 | 86.60 | 66.70 | 89.00 | 74.62 | 78.14 | 91.50 |
| + BM25 ($\alpha$) | 65.60 | 84.30 | 71.77 | 74.80 | 88.80 | 71.50 | 88.50 | 76.95 | 79.73 | 90.90 |
| + TF–IDF (RRF) | 57.20 | 82.20 | 65.52 | 69.54 | 87.00 | 64.60 | 84.10 | 71.55 | 74.62 | 89.80 |
| + BM25 (RRF) | 63.30 | 84.10 | 70.18 | 73.53 | 89.90 | 66.70 | 84.00 | 72.53 | 75.31 | 89.10 |
| 🤗 Qwen/Qwen3-Embedding-0.6B | | | | | | | | | | |
| Dense | 55.00 | 76.30 | 61.97 | 65.43 | 81.40 | 77.50 | 93.00 | 82.75 | 85.23 | 95.40 |
| + TF–IDF ($\alpha$) | 60.60 | 83.70 | 68.34 | 72.04 | 87.30 | 82.10 | 94.40 | 86.56 | 88.48 | 96.40 |
| + BM25 ($\alpha$) | 64.60 | 85.20 | 71.44 | 74.77 | 89.80 | 82.80 | 94.80 | 87.12 | 89.00 | 96.60 |
| + TF–IDF (RRF) | 57.50 | 82.50 | 65.92 | 69.92 | 88.40 | 77.30 | 93.50 | 83.26 | 85.77 | 95.50 |
| + BM25 (RRF) | 62.70 | 85.00 | 70.16 | 73.75 | 90.10 | 79.30 | 93.30 | 84.28 | 86.49 | 95.40 |

Table 16: Retrieval results on **Lifestyle & Reviews Domain** (VlogQA and ViRe4MRC).

| Method | VlogQA | | | | | ViRe4MRC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| TF–IDF | 13.40 | 34.60 | 19.55 | 23.10 | 47.00 | 3.70 | 17.50 | 7.24 | 9.63 | 23.70 |
| BM25 | 18.00 | 39.50 | 23.90 | 27.57 | 45.50 | 6.60 | 20.40 | 10.25 | 12.62 | 26.70 |
| ColBERT | 5.30 | 15.40 | 8.01 | 9.74 | 19.40 | 4.80 | 15.90 | 7.66 | 9.59 | 21.40 |
| SPLADE | 2.90 | 10.00 | 4.66 | 5.90 | 15.90 | 5.20 | 13.00 | 7.40 | 8.72 | 19.20 |
| text-embedding-3-large | | | | | | | | | | |
| Dense | 13.50 | 37.00 | 20.45 | 24.37 | 45.60 | 9.70 | 30.00 | 15.03 | 18.54 | 38.60 |
| + TF–IDF ($\alpha$) | 20.90 | 48.20 | 28.95 | 33.53 | 56.70 | 10.80 | 29.60 | 16.20 | 19.37 | 39.10 |
| + BM25 ($\alpha$) | 21.50 | 48.70 | 29.33 | 33.92 | 59.20 | 12.10 | 30.00 | 16.88 | 19.95 | 39.00 |
| + TF–IDF (RRF) | 20.10 | 48.40 | 28.72 | 33.42 | 58.70 | 10.70 | 26.90 | 15.32 | 18.05 | 37.20 |
| + BM25 (RRF) | 22.10 | 50.10 | 30.59 | 35.25 | 61.70 | 10.90 | 28.50 | 15.69 | 18.70 | 38.20 |
| 🤗 AITeamVN/Vietnamese_Embedding_v2 | | | | | | | | | | |
| Dense | 22.20 | 49.00 | 29.86 | 34.39 | 57.50 | 10.60 | 28.60 | 15.48 | 18.56 | 38.70 |
| + TF–IDF ($\alpha$) | 25.60 | 55.80 | 34.98 | 39.97 | 63.40 | 12.20 | 30.60 | 17.40 | 20.51 | 40.00 |
| + BM25 ($\alpha$) | 29.20 | 57.40 | 37.73 | 42.42 | 65.90 | **13.00** | 30.80 | **18.21** | **21.19** | 39.30 |
| + TF–IDF (RRF) | 24.10 | 55.20 | 33.79 | 38.91 | 63.10 | 12.20 | 29.40 | 16.98 | 19.91 | 37.10 |
| + BM25 (RRF) | 27.30 | 56.30 | 35.88 | 40.72 | 64.90 | **13.00** | 29.40 | 17.56 | 20.36 | 38.30 |
| 🤗 bkai-foundation-models/vietnamese-bi-encoder | | | | | | | | | | |
| Dense | 13.90 | 34.30 | 19.46 | 22.95 | 42.80 | 8.60 | 21.10 | 11.97 | 14.11 | 29.00 |
| + TF–IDF ($\alpha$) | 22.50 | 47.40 | 29.60 | 33.79 | 54.90 | 9.80 | 25.80 | 14.48 | 17.16 | 33.40 |
| + BM25 ($\alpha$) | 21.40 | 46.70 | 28.71 | 32.98 | 54.60 | 10.10 | 27.30 | 15.16 | 18.04 | 33.80 |
| + TF–IDF (RRF) | 19.90 | 48.90 | 28.26 | 33.15 | 58.50 | 9.40 | 25.40 | 14.14 | 16.82 | 32.50 |
| + BM25 (RRF) | 21.10 | 49.40 | 29.60 | 34.32 | 60.30 | 9.80 | 26.30 | 14.47 | 17.26 | 33.70 |
| 🤗 dangvantuan/vietnamese-document-embedding | | | | | | | | | | |
| Dense | 22.70 | 46.90 | 29.62 | 33.71 | 56.20 | 10.70 | 27.40 | 15.02 | 17.90 | 35.70 |
| + TF–IDF ($\alpha$) | 25.50 | 54.90 | 34.33 | 39.24 | 65.40 | 10.80 | 28.90 | 15.83 | 18.90 | 37.80 |
| + BM25 ($\alpha$) | 28.80 | 57.50 | 36.96 | 41.81 | 66.90 | 11.60 | 29.70 | 16.55 | 19.63 | 38.20 |

Continued on next page

Table 16 – continued from previous page

| Method | VlogQA | | | | | ViRe4MRC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| + TF–IDF (RRF) | 22.80 | 55.80 | 32.32 | 37.88 | 65.30 | 10.60 | 28.10 | 15.15 | 18.18 | 35.70 |
| + BM25 (RRF) | 25.60 | 57.20 | 35.03 | 40.32 | 66.00 | 11.00 | 28.20 | 15.55 | 18.50 | 36.90 |
| 🤗 sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | | | | | | | | | | |
| Dense | 4.50 | 14.80 | 7.16 | 8.95 | 20.50 | 4.70 | 16.10 | 7.30 | 9.33 | 22.10 |
| + TF–IDF ($\alpha$) | 14.40 | 33.20 | 19.67 | 22.86 | 41.30 | 8.40 | 22.20 | 11.89 | 14.30 | 29.90 |
| + BM25 ($\alpha$) | 12.00 | 32.30 | 17.65 | 21.11 | 39.80 | 8.50 | 23.20 | 12.48 | 15.00 | 31.10 |
| + TF–IDF (RRF) | 12.30 | 34.00 | 18.02 | 21.76 | 46.70 | 6.90 | 21.00 | 10.60 | 13.03 | 29.80 |
| + BM25 (RRF) | 13.10 | 37.90 | 19.66 | 23.93 | 49.10 | 7.30 | 21.90 | 11.26 | 13.75 | 31.20 |
| 🤗 intfloat/multilingual-e5-large | | | | | | | | | | |
| Dense | 23.90 | 49.20 | 31.09 | 35.38 | 57.00 | 11.80 | 29.00 | 16.77 | 19.66 | 37.00 |
| + TF–IDF ($\alpha$) | 29.00 | 59.90 | 37.85 | 43.06 | 67.90 | 12.00 | 30.40 | 17.02 | 20.17 | 37.60 |
| + BM25 ($\alpha$) | 29.00 | 57.90 | 37.79 | 42.56 | 66.80 | 12.40 | **31.60** | 17.84 | 21.10 | 39.40 |
| + TF–IDF (RRF) | 26.80 | 58.40 | 35.63 | 41.00 | 69.00 | 11.20 | 29.20 | 16.19 | 19.26 | 37.60 |
| + BM25 (RRF) | 28.40 | 56.90 | 36.69 | 41.48 | 67.80 | 11.90 | 30.20 | 16.94 | 20.07 | 38.30 |
| 🤗 jinaai/jina-embeddings-v3 | | | | | | | | | | |
| Dense | 24.00 | 51.90 | 32.03 | 36.74 | 59.80 | 9.90 | 27.50 | 14.73 | 17.74 | 36.00 |
| + TF–IDF ($\alpha$) | 28.20 | 60.60 | 37.73 | 43.16 | 68.40 | 12.10 | 29.70 | 16.90 | 19.90 | 39.10 |
| + BM25 ($\alpha$) | 30.00 | 61.40 | **39.53** | **44.73** | 68.30 | **13.00** | 30.70 | 17.82 | 20.84 | 40.10 |
| + TF–IDF (RRF) | 27.20 | 60.50 | 36.62 | 42.24 | 69.80 | 12.10 | 28.90 | 16.50 | 19.40 | 37.40 |
| + BM25 (RRF) | **30.70** | 60.60 | 39.41 | 44.43 | **69.90** | 12.70 | 30.50 | 17.10 | 20.20 | 39.20 |
| 🤗 BAAI/bge-m3 | | | | | | | | | | |
| Dense | 24.20 | 51.90 | 32.49 | 37.10 | 59.90 | 12.40 | 30.80 | 17.25 | 20.42 | 40.00 |
| + TF–IDF ($\alpha$) | 27.20 | 57.70 | 36.19 | 41.30 | 64.90 | 12.00 | 30.90 | 17.64 | 20.79 | 39.50 |
| + BM25 ($\alpha$) | 30.50 | 59.20 | 39.20 | 43.97 | 66.10 | 12.40 | 31.10 | 18.02 | 21.12 | **40.80** |
| + TF–IDF (RRF) | 25.10 | 56.90 | 34.62 | 39.92 | 64.50 | 10.30 | 28.70 | 15.86 | 18.91 | 38.00 |
| + BM25 (RRF) | 27.50 | 57.00 | 36.31 | 41.24 | 65.20 | 11.50 | 29.30 | 16.88 | 19.82 | 39.30 |
| 🤗 Snowflake/snowflake-arctic-embed-l-v2.0 | | | | | | | | | | |
| Dense | 20.70 | 49.60 | 28.91 | 33.81 | 59.30 | 9.20 | 24.80 | 13.66 | 16.29 | 33.10 |
| + TF–IDF ($\alpha$) | 26.70 | 59.00 | 36.63 | 41.96 | 67.30 | 10.60 | 28.20 | 15.66 | 18.63 | 35.50 |
| + BM25 ($\alpha$) | 29.40 | **61.80** | 38.91 | 44.34 | 69.60 | 11.50 | 28.30 | 16.37 | 19.21 | 36.60 |
| + TF–IDF (RRF) | 25.20 | 58.90 | 35.24 | 40.86 | 67.70 | 10.30 | 27.20 | 15.03 | 17.90 | 34.60 |
| + BM25 (RRF) | 27.60 | 58.30 | 36.75 | 41.88 | 68.80 | 11.30 | 27.40 | 15.71 | 18.47 | 36.40 |
| 🤗 Alibaba-NLP/gte-multilingual-base | | | | | | | | | | |
| Dense | 20.20 | 46.70 | 28.12 | 32.53 | 55.00 | 9.00 | 27.00 | 13.74 | 16.83 | 33.90 |
| + TF–IDF ($\alpha$) | 27.50 | 56.70 | 36.01 | 40.91 | 66.40 | 10.80 | 29.40 | 16.11 | 19.24 | 37.50 |
| + BM25 ($\alpha$) | 28.40 | 57.20 | 36.82 | 41.65 | 65.60 | 11.50 | 29.90 | 16.64 | 19.76 | 38.60 |
| + TF–IDF (RRF) | 24.40 | 56.00 | 33.65 | 38.94 | 67.80 | 9.70 | 28.40 | 14.92 | 18.09 | 36.90 |
| + BM25 (RRF) | 27.10 | 56.40 | 35.81 | 40.72 | 67.20 | 10.30 | 29.20 | 15.46 | 18.69 | 37.80 |
| 🤗 BAAI/bge-multilingual-gemma2 | | | | | | | | | | |
| Dense | 18.80 | 41.20 | 25.26 | 29.03 | 49.00 | 9.20 | 22.10 | 12.64 | 14.86 | 27.60 |
| + TF–IDF ($\alpha$) | 24.00 | 53.40 | 32.99 | 37.85 | 62.90 | 8.80 | 24.00 | 13.06 | 15.64 | 31.70 |
| + BM25 ($\alpha$) | 25.80 | 52.20 | 33.64 | 38.05 | 62.80 | 10.20 | 24.40 | 14.25 | 16.65 | 32.80 |
| + TF–IDF (RRF) | 22.70 | 51.80 | 31.12 | 36.00 | 63.50 | 8.70 | 23.30 | 12.51 | 15.04 | 31.90 |
| + BM25 (RRF) | 24.30 | 52.10 | 32.66 | 37.28 | 65.60 | 9.30 | 23.70 | 13.23 | 15.70 | 33.40 |
| 🤗 google/EmbeddingGemma-300m | | | | | | | | | | |
| Dense | 21.10 | 47.40 | 28.72 | 33.14 | 55.60 | 10.60 | 27.50 | 15.19 | 18.09 | 34.50 |
| + TF–IDF ($\alpha$) | 26.60 | 57.10 | 36.16 | 41.19 | 67.20 | 11.70 | 29.80 | 16.72 | 19.79 | 38.60 |
| + BM25 ($\alpha$) | 28.20 | 57.90 | 37.03 | 41.98 | 67.30 | 12.50 | 31.10 | 17.35 | 20.57 | 39.10 |
| + TF–IDF (RRF) | 25.10 | 56.70 | 34.22 | 39.55 | 69.00 | 10.70 | 28.00 | 15.56 | 18.50 | 37.30 |
| + BM25 (RRF) | 28.40 | 59.00 | 36.74 | 41.98 | 68.50 | 11.20 | 30.50 | 16.26 | 19.60 | 37.90 |
| 🤗 Alibaba-NLP/gte-Qwen2-1.5B-instruct | | | | | | | | | | |
| Dense | 3.60 | 11.20 | 5.40 | 6.75 | 16.40 | 2.10 | 9.80 | 4.07 | 5.40 | 13.50 |
| + TF–IDF ($\alpha$) | 13.80 | 34.50 | 19.62 | 23.12 | 44.70 | 4.00 | 16.10 | 7.32 | 9.39 | 22.40 |
| + BM25 ($\alpha$) | 15.20 | 31.70 | 19.83 | 22.64 | 40.20 | 5.50 | 17.30 | 8.83 | 10.84 | 22.80 |
| + TF–IDF (RRF) | 10.70 | 28.50 | 15.22 | 18.31 | 38.00 | 4.90 | 13.00 | 7.07 | 8.47 | 20.80 |
| + BM25 (RRF) | 10.30 | 26.80 | 14.79 | 17.61 | 38.60 | 5.60 | 14.10 | 7.80 | 9.27 | 21.80 |
| 🤗 Qwen/Qwen3-Embedding-0.6B | | | | | | | | | | |
| Dense | 23.50 | 48.10 | 30.71 | 34.82 | 55.80 | 11.50 | 28.10 | 15.97 | 18.81 | 35.00 |
| + TF–IDF ($\alpha$) | 27.20 | 56.60 | 36.11 | 40.99 | 67.10 | 11.20 | 31.00 | 16.62 | 19.99 | 39.00 |

Continued on next page

Table 16 – continued from previous page

| Method | VlogQA | | | | | ViRe4MRC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| + BM25 ($\alpha$) | 28.10 | 57.40 | 37.07 | 41.91 | 65.10 | 12.90 | 30.70 | 17.84 | 20.87 | 38.90 |
| + TF–IDF (RRF) | 24.70 | 57.40 | 34.36 | 39.81 | 67.40 | 10.10 | 29.00 | 15.22 | 18.46 | 37.10 |
| + BM25 (RRF) | 26.70 | 55.90 | 35.51 | 40.36 | 66.20 | 11.00 | 30.00 | 16.16 | 19.42 | 36.90 |

Table 17: Retrieval results on **Cross-domain Open Knowledge** (UIT-ViQuAD).

| Method | UIT-ViQuAD | | | | |
|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| TF–IDF | 50.00 | 91.00 | 64.57 | 71.05 | 94.00 |
| BM25 | 70.80 | 91.60 | 78.09 | 81.38 | 93.90 |
| ColBERT | 50.90 | 75.10 | 58.26 | 62.29 | 81.30 |
| SPLADE | 44.30 | 73.10 | 53.57 | 58.27 | 79.00 |
| text-embedding-3-large | | | | | |
| Dense | 71.20 | 92.40 | 78.75 | 82.09 | 96.00 |
| + TF–IDF ($\alpha$) | 79.30 | 97.20 | 86.24 | 88.96 | 98.20 |
| + BM25 ($\alpha$) | 82.90 | 97.00 | 88.46 | 90.58 | 98.70 |
| + TF–IDF (RRF) | 74.10 | 96.10 | 82.18 | 85.60 | 98.30 |
| + BM25 (RRF) | 77.80 | 95.80 | 84.59 | 87.37 | 98.70 |
| AITeamVN/Vietnamese_Embedding_v2 | | | | | |
| Dense | 82.20 | 98.10 | 88.24 | 90.67 | 99.00 |
| + TF–IDF ($\alpha$) | 84.40 | **99.30** | 90.18 | 92.45 | 99.30 |
| + BM25 ($\alpha$) | 89.10 | 99.20 | 93.00 | 94.53 | 99.30 |
| + TF–IDF (RRF) | 79.00 | 97.60 | 86.02 | 88.87 | 99.30 |
| + BM25 (RRF) | 82.10 | 97.50 | 87.95 | 90.32 | 99.10 |
| bkai-foundation-models/vietnamese-bi-encoder | | | | | |
| Dense | 68.00 | 88.40 | 74.62 | 77.94 | 92.10 |
| + TF–IDF ($\alpha$) | 77.40 | 95.00 | 83.93 | 86.66 | 97.20 |
| + BM25 ($\alpha$) | 81.20 | 96.10 | 86.71 | 89.02 | 97.80 |
| + TF–IDF (RRF) | 72.20 | 94.70 | 80.16 | 83.71 | 97.80 |
| + BM25 (RRF) | 75.70 | 94.90 | 82.50 | 85.54 | 97.80 |
| dangvantuan/vietnamese-document-embedding | | | | | |
| Dense | 75.70 | 95.60 | 83.00 | 86.10 | 97.40 |
| + TF–IDF ($\alpha$) | 81.40 | 97.50 | 87.45 | 89.95 | 99.00 |
| + BM25 ($\alpha$) | 85.50 | 98.70 | 90.34 | 92.40 | 99.30 |
| + TF–IDF (RRF) | 76.40 | 97.10 | 83.82 | 87.07 | 98.90 |
| + BM25 (RRF) | 79.60 | 97.10 | 86.07 | 88.79 | 99.10 |
| sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 | | | | | |
| Dense | 55.90 | 81.30 | 64.15 | 68.28 | 87.70 |
| + TF–IDF ($\alpha$) | 70.20 | 93.90 | 78.74 | 82.46 | 96.60 |
| + BM25 ($\alpha$) | 75.80 | 95.10 | 82.76 | 85.78 | 97.50 |
| + TF–IDF (RRF) | 66.30 | 93.20 | 75.45 | 79.76 | 96.50 |
| + BM25 (RRF) | 69.00 | 93.60 | 77.59 | 81.49 | 97.30 |
| intfloat/multilingual-e5-large | | | | | |
| Dense | 83.60 | 97.90 | 89.29 | 91.43 | 98.90 |
| + TF–IDF ($\alpha$) | 85.70 | 99.00 | 90.83 | 92.86 | 99.30 |
| + BM25 ($\alpha$) | **89.70** | 98.90 | **93.62** | **94.96** | **99.50** |
| + TF–IDF (RRF) | 78.60 | 98.40 | 86.52 | 89.49 | 99.20 |
| + BM25 (RRF) | 86.50 | 98.70 | 91.27 | 93.12 | 99.40 |
| jinaai/jina-embeddings-v3 | | | | | |
| Dense | 72.20 | 93.20 | 79.54 | 82.87 | 95.60 |
| + TF–IDF ($\alpha$) | 79.70 | 97.80 | 86.36 | 89.18 | 98.40 |
| + BM25 ($\alpha$) | 84.60 | 98.00 | 89.77 | 91.81 | 98.60 |
| + TF–IDF (RRF) | 74.50 | 97.10 | 83.04 | 86.53 | 98.30 |
| + BM25 (RRF) | 80.40 | 97.60 | 87.26 | 89.84 | 98.90 |
| BAAI/bge-m3 | | | | | |
| Dense | 80.60 | 96.40 | 86.62 | 89.04 | 98.40 |

Continued on next page

Table 17 – continued from previous page

| Method | UIT-ViQuAD | | | | |
|---|---|---|---|---|---|
| | P@1 | R@10 | MRR@10 | nDCG@10 | R@20 |
| + TF–IDF ($\alpha$) | 83.20 | 98.70 | 89.48 | 91.78 | 99.20 |
| + BM25 ($\alpha$) | 88.30 | 99.00 | 92.41 | 94.04 | 99.30 |
| + TF–IDF (RRF) | 77.20 | 97.60 | 84.95 | 88.07 | 99.00 |
| + BM25 (RRF) | 80.80 | 97.30 | 87.15 | 89.67 | 99.00 |
| 🤗 Snowflake/snowflake-arctic-embed-l-v2.0 | | | | | |
| Dense | 75.40 | 94.70 | 82.31 | 85.33 | 96.70 |
| + TF–IDF ($\alpha$) | 79.80 | 97.20 | 86.41 | 89.08 | 98.10 |
| + BM25 ($\alpha$) | 84.50 | 97.70 | 89.71 | 91.70 | 98.60 |
| + TF–IDF (RRF) | 75.30 | 97.10 | 83.51 | 86.87 | 98.30 |
| + BM25 (RRF) | 80.70 | 97.00 | 87.10 | 89.57 | 98.60 |
| 🤗 Alibaba-NLP/gte-multilingual-base | | | | | |
| Dense | 75.00 | 95.20 | 82.34 | 85.49 | 97.00 |
| + TF–IDF ($\alpha$) | 80.40 | 98.20 | 86.92 | 89.69 | 98.90 |
| + BM25 ($\alpha$) | 85.00 | 98.50 | 90.26 | 92.31 | 99.20 |
| + TF–IDF (RRF) | 76.30 | 97.50 | 84.12 | 87.42 | 98.80 |
| + BM25 (RRF) | 82.30 | 98.10 | 88.43 | 90.83 | 99.10 |
| 🤗 BAAI/bge-multilingual-gemma2 | | | | | |
| Dense | 70.90 | 93.00 | 78.74 | 82.23 | 96.00 |
| + TF–IDF ($\alpha$) | 78.70 | 96.80 | 85.60 | 88.37 | 98.20 |
| + BM25 ($\alpha$) | 83.10 | 97.50 | 88.42 | 90.66 | 98.50 |
| + TF–IDF (RRF) | 75.50 | 95.60 | 82.95 | 86.07 | 98.30 |
| + BM25 (RRF) | 77.90 | 96.30 | 84.82 | 87.65 | 98.60 |
| 🤗 google/EmbeddingGemma-300m | | | | | |
| Dense | 76.90 | 94.90 | 83.45 | 86.26 | 97.40 |
| + TF–IDF ($\alpha$) | 81.80 | 98.50 | 88.43 | 90.94 | 98.80 |
| + BM25 ($\alpha$) | 86.20 | 98.50 | 91.24 | 93.06 | 99.10 |
| + TF–IDF (RRF) | 77.40 | 97.80 | 85.00 | 88.17 | 98.60 |
| + BM25 (RRF) | 83.50 | 98.20 | 89.30 | 91.52 | 99.00 |
| 🤗 Alibaba-NLP/gte-Qwen2-1.5B-instruct | | | | | |
| Dense | 56.70 | 84.10 | 65.64 | 70.09 | 88.50 |
| + TF–IDF ($\alpha$) | 72.30 | 94.90 | 80.73 | 84.22 | 97.10 |
| + BM25 ($\alpha$) | 79.20 | 95.50 | 85.30 | 87.82 | 96.90 |
| + TF–IDF (RRF) | 67.10 | 92.20 | 75.85 | 79.83 | 96.00 |
| + BM25 (RRF) | 69.30 | 93.10 | 77.90 | 81.61 | 96.20 |
| 🤗 Qwen/Qwen3-Embedding-0.6B | | | | | |
| Dense | 73.70 | 94.10 | 80.76 | 84.01 | 97.20 |
| + TF–IDF ($\alpha$) | 78.20 | 97.80 | 85.81 | 88.79 | 99.10 |
| + BM25 ($\alpha$) | 85.30 | 98.30 | 90.41 | 92.38 | 99.30 |
| + TF–IDF (RRF) | 74.10 | 97.60 | 82.96 | 86.58 | 98.70 |
| + BM25 (RRF) | 81.60 | 98.00 | 88.06 | 90.53 | 99.30 |