

Agentic Retrieval-Augmented Generation for Vietnamese Elementary History Education: A Hybrid Search, Cross-Encoder Reranking, and Pedagogical Assessment Framework

Hoang Ngoc Long¹, Nguyen Minh Duc², Trinh Thanh Do^{3*}

¹Reigate Grammar School Vietnam, Hanoi, Vietnam

Email: longhoang4309@gmail.com

²Techcom Life Insurance Joint Stock Company, Vietnam

Email: phucanhthien@gmail.com

³Rikai Technology, Danang, Vietnam

Email: Dotrinh.cs@gmail.com

*Corresponding author

Abstract: Providing accurate, engaging, and age-appropriate historical information to elementary students remains challenging because general-purpose large language models may generate unsupported facts, confuse historical periods, or use unsuitable language. This study presents Dai Viet Kids, a curriculum-grounded, agent-oriented retrieval-augmented generation framework for Vietnamese history education in Grades 4 and 5. The framework combines hierarchy-preserving breadcrumb chunking, sparse-dense hybrid retrieval with reciprocal rank fusion, cross-encoder reranking, intent routing, and persona-guided response generation with preliminary safety guardrails. We conducted a pilot evaluation using 46 curriculum-aligned questions and four adversarial robustness probes over 914 indexed textbook chunks. The full retrieval pipeline achieved Hit@1 of 74.00%, Hit@3 and Hit@5 of 94.00%, and a mean reciprocal rank at cutoff five (MRR@5) of 0.8233. BM25 alone achieved Hit@1 of 72.00%, Hit@5 of 94.00%, and an MRR@5 of 0.7827, so the difference between the two configurations is concentrated at Hit@3, with Hit@1 differing by a single query and Hit@5 identical; at this benchmark size no pairwise difference in Table 2 attains statistical significance, as Section 7 shows. Automated generation assessment yielded a faithfulness score of 85.05% and an answer-relevancy score of 87.94%. These findings provide preliminary evidence that hybrid retrieval and reranking can improve curriculum-grounded historical question answering. However, the small benchmark, limited robustness testing, and absence of teacher or learner evaluation mean that classroom effectiveness and child safety remain to be validated in future studies.

Keywords: Agentic RAG; Vietnamese History Education; Hybrid Search; Reciprocal Rank Fusion; Cross-Encoder Reranking; Hallucination Mitigation; RAGAS; Multi-Agent LLM Systems; EdTech

1. Introduction

Historical education plays a foundational role in fostering national identity, cultural awareness, and critical thinking in elementary school children. In Vietnam, the national curriculum for Grade 4 and Grade 5 introduces students to crucial historical milestones, spanning from the legendary Hong Bang Dynasty and King Hung, through the resistance wars of the Trung Sisters, Ngo Quyen's naval triumph on the Bach Dang River, to the 20th-century Dien Bien Phu campaign and the 1975 Spring Victory.

However, traditional teaching methodologies often rely on rote memorization of dates, names, and event chronologies, resulting in passive learning and reduced student engagement. While conversational Artificial Intelligence powered by Large Language Models (LLMs) offers an unprecedented opportunity to transform history learning into an interactive, inquiry-based dialogue, raw LLMs exhibit fatal limitations in primary education:

- **Factual Hallucinations & Chronological Anachronisms:** General-purpose LLMs frequently blend disparate historical eras (e.g., claiming tanks were used in ancient battles or confusing King Quang Trung with Nguyen Anh). Such fabrications are a well-documented failure mode of unconstrained neural text generation [1].

- **Lack of Age-Appropriate Pedagogy:** Standard LLM responses are either overly dry, verbose, or laden with academic jargon unsuitable for 9-11-year-old learners.
- **Child Safety and Content Vulnerabilities:** Without domain-specific input/output guardrails, educational chatbots risk answering off-topic prompts or generating unverified historical narratives.

To resolve these challenges, Retrieval-Augmented Generation (RAG) [2] grounds LLM generation in authoritative external corpora, a strategy shown to reduce hallucination substantially across knowledge-intensive natural-language-processing tasks [2], [3]. Nonetheless, naive RAG pipelines relying solely on dense vector cosine similarity struggle with Vietnamese historical terminology, proper nouns, dynasty names, and short, keyword-dense queries typical of young children a gap consistent with the weaker zero-shot generalization of dense bi-encoder retrievers reported on heterogeneous benchmarks, including entity-centric retrieval tasks.

In this work, we present an end-to-end Agentic RAG Framework for Vietnamese Elementary History Education developed within the Dai Viet Kids project. Our major scientific and technical contributions include:

- **Curriculum-Preserving Semantic Chunking:** A hierarchical Markdown parser that injects breadcrumb path metadata (Unit → Chapter → Section → Period)

into 914 chunks derived from national Grade 4 and Grade 5 history textbooks.

- Two-Stage Hybrid Search & Cross-Encoder Reranking: A hybrid retrieval engine combining lexical BM25 indexing with dense neural embeddings via Reciprocal Rank Fusion (RRF). The fused list serves as a high-recall candidate generator at depth $M = 15$, not as the final ranking; a Cross-Encoder reranking model (ms-marco-MiniLM-L-6-v2) then performs the fine-grained ordering over that pool. This division of labour is what allows a lexically strong but semantically shallow signal and a semantically broad but lexically brittle signal to be combined without either one dominating the final answer.
- Multi-Agent Persona & Safety Architecture: A decoupled orchestration architecture, aligned with recent LLM-based multi-agent design principles, comprising an Intent Router, Knowledge Retrieval Agent, Interactive Quiz Agent, and Roleplay Agent, governed by the friendly pedagogical persona of “Cù Rùa Thông Thái” and by programmable input/output guardrails [4], [5].
- Empirical Scientific Benchmark: Construction of a 50-query Golden Historical QA benchmark stratified by Bloom’s Taxonomy, and systematic ablation experiments quantifying the contribution of each component of our proposed architecture across IR (Hit@K, MRR@5) and generative (Faithfulness, Relevancy, Latency) dimensions.

2. Related Work

2.1 Retrieval-Augmented Generation (RAG) in Education

Retrieval-Augmented Generation has emerged as the state-of-the-art paradigm for knowledge-intensive NLP tasks [2], combining a non-parametric retriever with a parametric generator so that factual claims can be traced back to a verifiable source rather than to the model’s internal parameters alone. In educational contexts, RAG enables personalized tutoring systems by retrieving verified textbook passages before prompting the generative backbone [6]. Educational applications of RAG have since been surveyed systematically, spanning tutoring, assessment and content generation [7], which is particularly important given that hallucination remains one of the most persistent and well-catalogued failure modes of natural-language generation systems [1]. A broader systematic review of AI applications in higher education further finds that intelligent tutoring systems and adaptive personalization are among the most active application areas, yet notes a persistent gap between educator involvement and system design [8], underscoring the need for pedagogically grounded architectures such as the one proposed here. However, application of RAG to Vietnamese domain-specific education, particularly primary-school history, remains scarcely investigated; the present work also builds on Vietnamese-specific language-modeling resources such as PhoBERT [9], which established the first large-scale monolingual pretrained representations for Vietnamese and continues to inform embedding choices for Vietnamese information retrieval.

Surveys of the field distinguish three successive RAG paradigms: Naive RAG, which performs a single index-retrieve-generate pass; Advanced RAG, which adds pre- and post-retrieval optimisation such as query rewriting and re-ranking; and Modular RAG, in which retrieval, routing, and generation are separable, independently replaceable components [6]. The architecture presented here belongs to the third category: the Intent Router, the retrieval engine, the reranker, and the persona-bound synthesis agent are distinct modules with separate failure modes, which is what makes component-level ablation (Section 5.1) and component-level safety auditing (Section 7) possible at all. The original formulation of RAG frames this as a pairing of parametric memory, held in the generator’s weights, with non-parametric memory, held in an external index that can be inspected and replaced without retraining [2]; for a classroom deployment, the second half of that pairing is the property that matters, because a factual error in the textbook index can be corrected by editing a chunk rather than by fine-tuning a model. On the Vietnamese side, the resources available for grounding and evaluation remain thin: UIT-ViQuAD, the reference Vietnamese machine-reading-comprehension benchmark, comprises roughly 23,000 question-answer pairs over 174 Wikipedia articles [10], an encyclopedic register quite unlike the simplified, narrative prose of a Grade 4 history textbook. This register gap is part of why we construct a curriculum-specific benchmark rather than reusing an existing one.

2.2 Hybrid Retrieval and Neural Reranking

While dense bi-encoders such as Dense Passage Retrieval (DPR) [3] and Sentence-BERT [11] effectively capture broad semantic similarity by mapping queries and passages into a shared embedding space, they frequently underperform on exact keyword matching, dates, and historical proper nouns (e.g., “Hịch tướng sĩ”, “Đại La”, “Nơ-trang Long”) — a weakness documented across heterogeneous retrieval benchmarks [12]. Sparse retrieval models such as BM25, grounded in the classical probabilistic relevance framework [13], excel at exact lexical matches and remain a strong baseline even against learned dense retrievers, a robustness confirmed across the 18 datasets of the BEIR zero-shot benchmark.⁹ Hybrid search combining dense and sparse retrievers via Reciprocal Rank Fusion (RRF) has been shown to produce more robust candidate lists than either signal alone [14]. To further eliminate false positives, a Cross-Encoder is applied to compute full cross-attention between query and candidate passages, at the cost of higher inference latency but with a substantial gain in ranking precision [15], [16].

The quantitative evidence for this two-stage design is direct. In the BEIR zero-shot evaluation of nine retrieval architectures over eighteen datasets, BM25 proved the strongest single baseline, and the only configurations that surpassed it on more than half of the datasets were reranking-based: BM25 followed by a cross-encoder outperformed BM25 alone on 11 of 17 zero-shot tasks, at the cost of substantially slower retrieval [12]. The same study reports that dense retrievers surpass BM25 only where the evaluation domain overlaps the data they were trained on, and underperform it otherwise. Our own pipeline is precisely the BM25-plus-cross-encoder configuration that BEIR

identifies as the most robust, applied to a corpus with no overlap with any public training set. Two caveats attach to this transfer of evidence: BEIR is restricted to English tasks, and its documents average a few hundred words, so its numbers motivate our design rather than predicting our results. Comparable reasoning applies to the fusion stage. When DPR combines sparse and dense retrieval, it does so by rescoring the union of the top-2000 passages from each retriever with $\text{BM25}(q,p) + \lambda \cdot \text{sim}(q,p)$ at $\lambda = 1.1$, rather than by merging short ranked lists [3]; fusion is treated throughout the literature as a recall-stage operation performed over deep candidate pools. Two families of alternatives to this sparse-plus-cross-encoder arrangement deserve mention, since we did not adopt either. Late-interaction models such as ColBERT retain a per-token embedding for every document and score by summing maximum-similarity matches, recovering much of the cross-encoder’s accuracy at a fraction of its query-time cost [17]; learned sparse models such as SPLADE keep an inverted index but replace term frequencies with learned, vocabulary-projected weights, giving lexical expansion without a dense index [18]. Both would require Vietnamese-language training data we do not have, and both are surveyed alongside cross-encoder reranking in the standard reference on transformer-based ranking [19]. Their absence is a limitation of scope, not evidence that the chosen design is optimal.

2.3 Agentic and Multi-Agent LLM Systems for Safe Deployment

Beyond single-pass retrieval, recent work has moved LLM-based systems from monolithic prompting toward agentic, multi-agent architectures in which specialized components handle planning, tool use, and role-specific generation [4]. Such decompositions are particularly suited to child-facing educational assistants, where an intent router, a knowledge-grounded synthesis agent, and interaction-specific agents (e.g., quiz or roleplay) can be independently audited and constrained. Complementing this architectural shift, toolkits for programmable guardrails allow developers to attach interpretable, user-defined dialogue rails on top of an LLM, independent of the underlying model, to prevent off-topic, unsafe, or ungrounded responses [5]. We adopt this programmable-guardrail philosophy for the Child Safety Guardrail and output sanitization stages of our pipeline (Section 3.4). Looking beyond passage-level retrieval, graph-structured extensions of RAG that summarize entity relationships across an entire corpus [20] point toward a natural future direction for modeling cross-dynasty relationships in Vietnamese history, which we revisit in Section 7.

Surveys of LLM-based multi-agent systems organise this design space along four axes: the agents-environment interface, agent profiling, agent communication, and capability acquisition [4]. Our framework instantiates the first three explicitly. The environment interface is the retrieval index; profiling is carried out through the fixed pedagogical persona of “Cụ Rùa Thông Thái”; and communication is deliberately constrained to a star topology through the Intent Router, since free inter-agent messaging would make the provenance of a claim shown to a child difficult to audit. Capability acquisition is static in our

system, as no agent learns from interaction. On the guardrail side, NeMo Guardrails implements its rails as a runtime proxy interposed between user and model, with dialogue flows written in a dedicated modelling language, Colang, and interpreted by a dialogue manager at inference time [5]. The significance of that architecture for child-facing deployment is that the rails are external to the model and therefore inspectable and editable by an educator, in contrast to safety behaviour baked in during alignment training. Our Intent Router is deliberately simpler than the interleaved reason-then-act loop that has become standard for LLM agents [21]: it classifies once and dispatches, with no iterative tool use, because an unbounded loop is difficult to bound in latency and difficult to explain to a teacher reviewing a transcript. For the same reason we do not currently implement self-reflective retrieval, in which the model emits critique tokens deciding whether to retrieve again and whether its own draft is supported by the evidence [22]; that mechanism is an attractive route to raising the Hallucination Defense Rate and we return to it in Section 7.

2.4 LLM Evaluation and the RAGAS Framework

Automated evaluation of RAG systems without costly human annotation has made substantial progress through the RAGAS (Retrieval Augmented Generation Assessment) framework [23]. RAGAS decouples evaluation into retrieval metrics (Context Precision, Context Recall) and generation metrics (Faithfulness, Answer Relevancy). We adapt these foundational metrics, together with an adversarial Hallucination Defense Rate inspired by the broader NLG hallucination literature [1], to assess factual consistency and pedagogical suitability in elementary Vietnamese education.

Concretely, RAGAS estimates faithfulness by prompting an evaluator LLM to decompose an answer into a set S of atomic statements, verifying each statement against the retrieved context, and reporting $F = |V| / |S|$, where V is the subset judged entailed; answer relevancy is estimated by generating candidate questions from the answer and measuring their similarity to the original query. In the original evaluation, conducted with gpt-3.5-turbo-16k on the WikiEval dataset, these estimators agreed with human pairwise judgements in around 95% of cases for faithfulness [23]. The hallucination literature further separates intrinsic hallucination, where the output contradicts the provided source, from extrinsic hallucination, where the output cannot be verified against it [1]. The distinction matters for how our two-generation metrics should be read: the Faithfulness score penalises both failure modes, whereas the Hallucination Defense Rate defined in Section 4.2 measures something narrower, namely whether the agent actively contradicts a false premise supplied by the student. A system could in principle score well on faithfulness by abstaining and still score poorly on HDR.

3. System Architecture and Methodology

The Dai Viet Kids Agentic RAG framework is organized as a pipeline of six cooperating stages: input safety filtering, intent routing, dual-stage hybrid retrieval, cross-encoder reranking, persona-driven multi-agent synthesis, and output sanitization. Figure 1 depicts the complete end-to-end architecture, including

the routing boundary between knowledge-grounded questions and specialized interactive agents.

Figure 1: End-to-End Architecture of the Dai Viet Kids Agentic RAG Framework

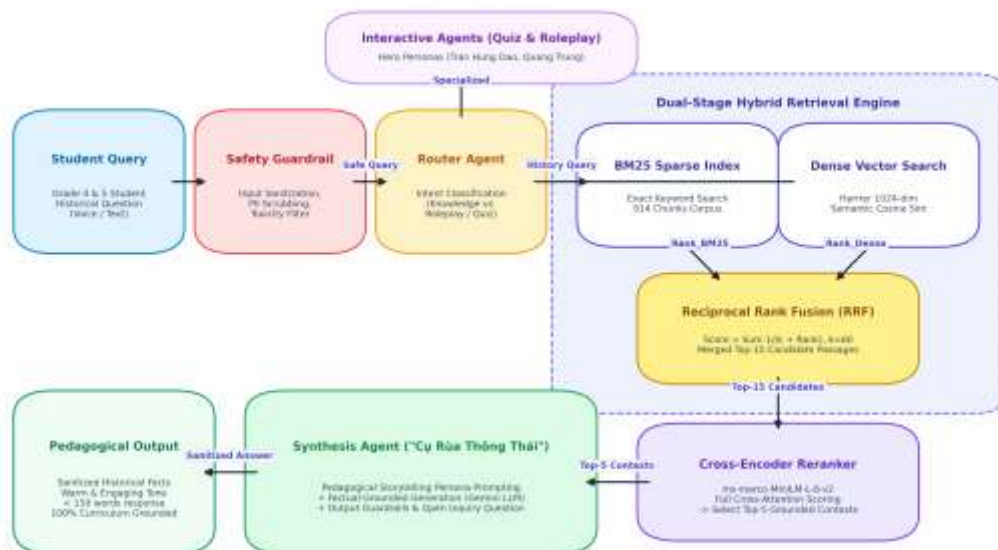


Figure 1. End-to-end architecture of the Dai Viet Kids Agentic RAG framework, illustrating the pipeline from query sanitization and intent routing to dual-stage hybrid retrieval and pedagogical synthesis.

3.1 Structural Breadcrumb Semantic Chunking

Raw Grade 4 and Grade 5 history textbooks contain intricate hierarchical structures (Units, Chapters, Historical Epochs). Standard fixed-window chunking destroys thematic context when a paragraph is separated from its chapter title, a problem that is especially acute for children’s queries, which tend to be short and rely heavily on surrounding context to disambiguate proper nouns.

Our SemanticChunker parses markdown headers (#, ##, ###) and constructs a dynamic breadcrumb prefix prepended to each

chunk, so that every indexed unit carries its own hierarchical provenance: $\text{Chunk} = [\text{Unit} > \text{Chapter} > \text{Section}] + \text{Body Paragraph}$. Across the entire Grade 4 and Grade 5 curriculum, this yields 914 structured chunks indexed jointly in ChromaDB (for dense retrieval) and in the BM25 corpus (for lexical retrieval). Figure 2 illustrates this transformation on a representative passage about the founding of the Van Lang Kingdom.

Figure 2: Structural Breadcrumb Semantic Chunking Preserving Grade 4 & 5 Hierarchy

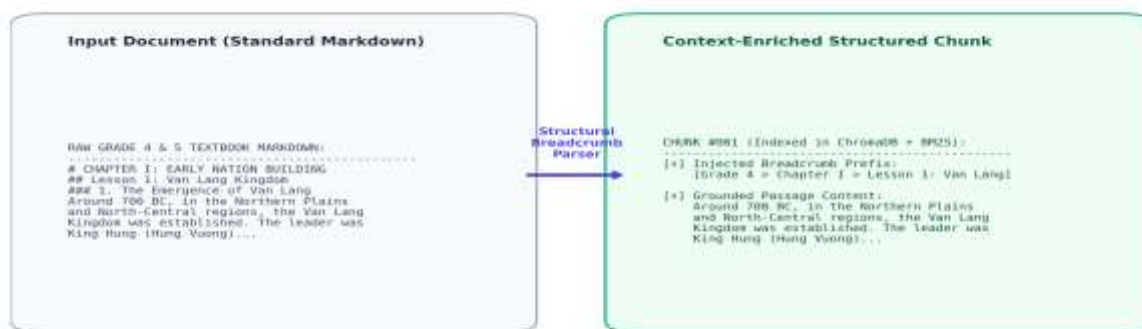


Figure 2. Structural Breadcrumb Semantic Chunking. The parser converts raw textbook markdown into context-enriched chunks by injecting hierarchical prefixes, preserving the curriculum structure for retrieval indices.

3.2 Hybrid Retrieval with Reciprocal Rank Fusion (RRF)

Given query q , the lexical retriever computes BM25 scores over chunk vocabulary D following the probabilistic relevance framework [13], while the vector retriever computes cosine distance over 1024-dimensional dense embeddings generated by our Harrier embedding microservice, in the spirit of dual-

encoder dense retrieval [3], [11]. Retrieval proceeds in three steps:

One Vietnamese-specific consideration governs the lexical arm. Vietnamese uses white space to separate syllables rather than words, and roughly 85% of Vietnamese word types consist of two or more syllables [9], so a BM25 index built on whitespace

tokenisation fragments exactly the multi-syllable proper nouns that carry the discriminative signal in this corpus: “Ngô Quyền”, “Bạch Đằng” and “Hịch tướng sĩ” become two or three independent terms, each individually common across the curriculum. This fragmentation is the mechanism by which purely lexical retrieval loses precision on entity-dense queries, and it is one of the reasons the dense arm is retained in the fusion despite its weak standalone performance: the two retrievers may fail on different query shapes, although the present ablation does not measure their per-query overlap and therefore does not establish this. Word-segmented indexing, using an established Vietnamese segmenter such as VnCoreNLP, is a further avenue we have not yet evaluated and which we flag in Section 7. On the dense side, the corresponding question is whether the embedding space was ever aligned to Vietnamese at all; multilingual sentence embeddings are typically obtained by distilling a monolingual English teacher into a multilingual student on parallel data, and the quality of the Vietnamese portion of that alignment bounds what the dense arm can contribute [24], [25].

Sparse Ranking: $R_BM25(q) = \text{rank}(BM25(q, D))$

Dense Ranking: $R_Dense(q) = \text{rank}(\cos(e_q, e_d))$

Reciprocal Rank Fusion: $RRF(d) = 1 / (k + r_BM25(d)) + 1 / (k + r_Dense(d))$, $k = 60$

where $k = 60$ is the smoothing constant used in the original RRF formulation, where it serves to damp the influence of high ranks contributed by outlier systems [14]. The top $M = 15$ candidate documents with highest $RRF(d)$ are routed to the reranker. $M = 15$ is therefore the operating point of the fusion stage: the quantity that matters for the pipeline is the recall of the fused list at depth 15, not its precision at depth 5. The same design logic underlies the hybrid retriever of DPR, which fuses the union of the top-2000 BM25 and dense lists before re-scoring rather than fusing short lists directly [3].

3.3 Cross-Encoder Neural Reranking

Bi-encoder embeddings compress passage semantics into fixed-length vectors, losing fine-grained cross-token interactions between query and passage. We deploy a Cross-Encoder (cross-encoder/ms-marco-MiniLM-L-6-v2) that passes the concatenated sequence [CLS] Query [SEP] Passage [SEP] through full Transformer attention layers, following the BERT cross-encoder re-ranking formulation of [16]:

Table 1. Distribution of the Golden Benchmark QA Dataset ($N = 50$). The first three strata follow the revised Bloom taxonomy; the Adversarial / Out-of-Domain stratum is a robustness probe outside the Bloom hierarchy.

Cognitive Level (Bloom-aligned)	Description / Purpose	Count	Proportion
Recall	Factual recall of dates, locations, figures, and treaty terms	35	70.0%
Comprehension	Explaining significance, tactical motives, and event causes	7	14.0%
Analysis / Reasoning	Comparative analysis between historical epochs and strategies	4	8.0%
Adversarial / Out-of-Domain (non-Bloom robustness probe)	Trap questions testing chronological and counterfactual resistance	4	8.0%

$$s(q, d) = \sigma(W \cdot \text{Transformer}([q; d])_{[\text{CLS}]})$$

The specific checkpoint is a six-layer MiniLM, a transformer compressed by deep self-attention distillation [26], fine-tuned on MS MARCO, a corpus of 1,010,916 anonymised Bing search queries over 8.8 million web passages [27]. Both the architecture and the supervision are therefore English web-search artefacts, which is the provenance behind the reranker limitation recorded in Section 7. The top $K = 5$ passages sorted by $s(q, d)$ form the final grounded context $C = \{c_1, c_2, \dots, c_K\}$.

3.4 Multi-Agent Synthesis and Child-Safety Guardrails

The SynthesisAgent coordinates generation under pedagogical constraints, drawing on the programmable-rails philosophy of controllable LLM applications [5] and on multi-agent orchestration principles [4] to keep each responsibility auditable and independently testable:

- **ChildSafetyGuardrail:** Pre-execution filter scrubbing toxic keywords, inappropriate topics, and personally identifiable information (PII).
- **Persona Prompting (“Cụ Rùa Thông Thái”):** Directs the LLM to address students as “cháu” or “nhà sử học nhí”, maintain an encouraging, storytelling tone, restrict length to under 150 words, and conclude with an open inquiry question. The prompt also elicits a brief intermediate reasoning step before the final answer, following the chain-of-thought prompting pattern [28], which we use here to keep chronological ordering explicit rather than to solve arithmetic.
- **Abstention & Grounded Fallback:** When retrieved context is sparse (under 30 tokens) or when API rate limits occur, the system safely falls back to sanitized verified textbook excerpts rather than hallucinating.

4. Experimental Setup and Golden Benchmark

4.1 Benchmark Dataset Construction

We constructed a specialized evaluation dataset comprising 50 Golden QA Pairs carefully authored from official Grade 4 and Grade 5 history textbooks. The benchmark is balanced across grade levels (25 Grade 4, 25 Grade 5) and categorized into three Bloom-aligned cognitive levels plus a fourth, non-Bloom adversarial / out-of-domain stratum, following the revised taxonomy [29], [30], which remains a standard framework for stratifying question difficulty by cognitive demand in educational assessment design.

Cognitive Level (Bloom-aligned)	Description / Purpose	Count	Proportion
Total	Comprehensive elementary history coverage	50	100.0%

4.2 Evaluation Metrics

- Hit Rate@K (Hit@K): the proportion of test queries for which at least one ground-truth context chunk appears within the top-K retrieved passages.
- Mean Reciprocal Rank at cutoff five (MRR@5): the average of the reciprocal of the rank at which the first relevant document is retrieved within the top five results, with queries whose first relevant document falls outside the top five assigned a reciprocal rank of zero; formally, $MRR@5 = (1/N) \sum 1/r_i$ for $r_i \leq 5$ and 0 otherwise. The metric therefore penalises systems that bury the correct passage lower in the ranked list, and is bounded below by the corresponding Hit@5 divided by five.
- Faithfulness Score (F): following the RAGAS protocol, the fraction of atomic claims in the synthesized answer that are verifiably grounded in the retrieved context [23].

- Answer Relevancy Score (R): the semantic alignment between the student’s question and the generated answer, computed with the RAGAS answer-relevancy procedure [23].
- Hallucination Defense Rate (HDR): the percentage of adversarial or counterfactual queries for which the agent successfully refutes a false premise or identifies a factual error, in line with hallucination-robustness evaluation practice in the NLG literature [1].
- Retrieval and Generation Latency: measured in milliseconds (ms) and seconds (s) per query, capturing the operational feasibility of real-time classroom deployment.

5. Results and Discussion

5.1 Retrieval Ablation Study

We conducted an ablation experiment comparing four retrieval configurations over all 50 benchmark queries against the 914-chunk corpus.

Table 2. Quantitative ablation results across retrieval architectures (N = 50). Bracketed intervals are 95% Wilson confidence intervals for a binomial proportion with n = 50. The latency column reports the retrieval stage only, that is query embedding, sparse and dense lookup, reciprocal rank fusion and, for the proposed configuration, cross-encoder reranking; it excludes answer generation, which is reported separately in Table 3. Differences between configurations are not statistically significant at this sample size; see Section 7.

Retrieval Architecture	Hit@1 (%) [95% CI]	Hit@3 (%) [95% CI]	Hit@5 (%) [95% CI]	MRR@5	Avg Retrieval Latency (ms)
1. Vector Only (Dense Embeddings)	4.00 [1.1–13.5]	8.00 [3.2–18.8]	8.00 [3.2–18.8]	0.0533	722.19
2. BM25 Only (Sparse Lexical)	72.00 [58.3–82.5]	86.00 [73.8–93.0]	94.00 [83.8–97.9]	0.7827	7.47
3. Hybrid Search (BM25 + Dense RRF)	72.00 [58.3–82.5]	86.00 [73.8–93.0]	86.00 [73.8–93.0]	0.7900	722.24
4. Proposed (Hybrid + Cross-Encoder Reranker)	74.00 [60.4–84.1]	94.00 [83.8–97.9]	94.00 [83.8–97.9]	0.8233	1863.30

Figure 3: Retrieval Performance Across 4 Ablation Configurations (N = 50 Queries)

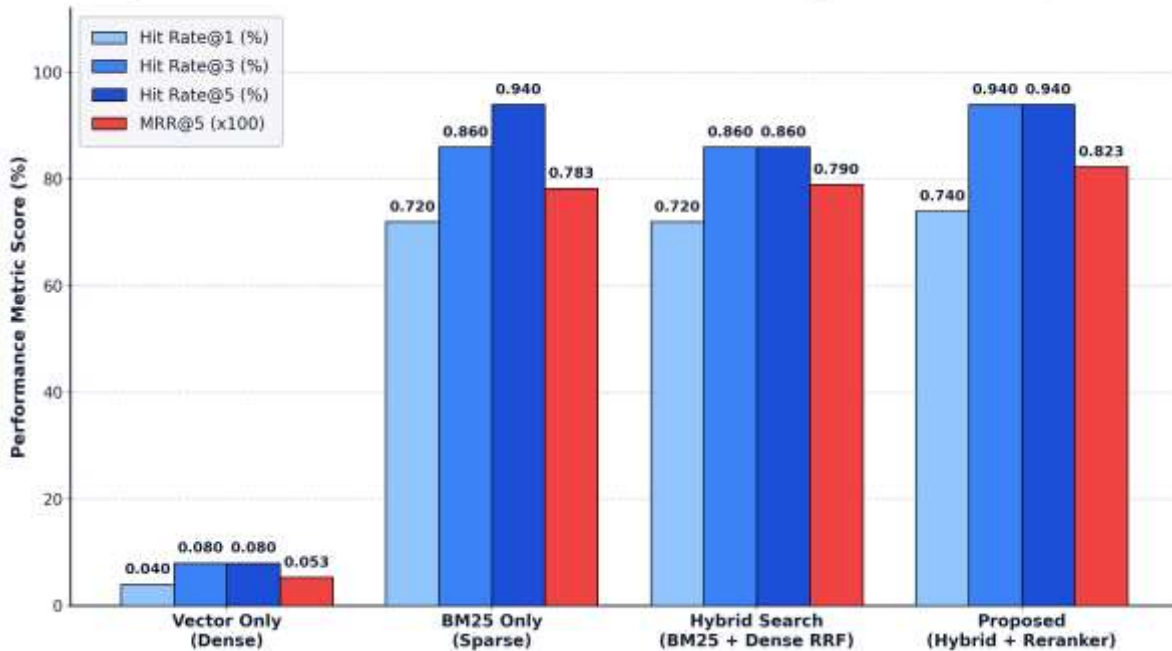


Figure 3. Retrieval performance metrics across the four ablation configurations (N = 50 queries), highlighting the overall improvement of the proposed Hybrid + Reranker pipeline.

Key Retrieval Insights:

- **Dense Vector Limitations in the Historical Domain:** Dense vector search alone achieved only a 4.0% Hit@1 and 0.0533 MRR@5. Vietnamese historical text contains high-density named entities (e.g., “Trịnh - Nguyễn phân tranh”, “Chiến dịch Việt Bắc thu - đông 1947”), where generalized sentence embeddings fail to distinguish subtle epoch distinctions without exact keyword signals, consistent with the BEIR finding that dense retrievers exceed BM25 only under high overlap between their training distribution and the target domain, and fall behind it otherwise. Our embedding service was not trained on Vietnamese curricular history, so the conditions under which dense retrieval is reported to win are absent here; the result should be read as evidence about this deployment, not as a general claim about dense retrieval [12].

- **Complementary Power of Hybrid RRF:** BM25 demonstrates strong lexical recall (72.0% Hit@1). Fusing it with dense search via RRF raises MRR@5 from 0.7827 to 0.7900 while holding Hit@1 and Hit@3 constant, i.e. the fusion improves the average position of the first relevant chunk. At Hit@5, however, the fused list scores 86.00% against BM25’s 94.00%. This is not an anomaly but the expected behaviour of rank fusion when the two input lists are of very unequal quality: RRF was originally validated on TREC runs of broadly comparable strength, where it exceeded the best individual system by 4–5% [14], whereas here it fuses a 0.7827-MRR@5 list with a 0.0533-MRR@5 list. The $k = 60$ constant damps, but does not eliminate, the ability of the weaker list to displace a borderline BM25 hit out of the top five. Two consequences follow, and they are the reason the fusion stage is deployed at $M = 15$ rather than at $K = 5$. First, the shallow-cutoff penalty disappears at the depth the system actually uses: because the

Cross-Encoder can only re-order documents that the fused list already contains, the 94.00% Hit@5 of the full pipeline establishes a lower bound of $\text{Recall}@15 \geq 94.00\%$ for the fused candidate pool. Second, the dense arm may contribute complementary candidates to the fused pool even at 4.0% standalone Hit@1, but its independent contribution is not established by the present ablation: Table 2 reports no per-query cross-tabulation of sparse and dense successes, and the two case studies of Section 5.3 are illustrative rather than evidential. Establishing it would require reporting, over the same 50 queries, the counts of queries retrieved only by BM25, only by the dense arm, by both, and by neither. We therefore report hybrid retrieval as a recall-oriented candidate generator whose contribution is realised through the reranker, and not as a standalone ranker that supersedes BM25 at every cutoff.

- **Reranker Efficacy:** The Cross-Encoder raises Hit@3 from 86.00% under Hybrid RRF to 94.00% and MRR@5 from 0.7900 to 0.8233. Because the reranker can only promote documents that the fusion stage has already retrieved, this also establishes a lower bound of $\text{Recall}@15 \geq 94.00\%$ for the fused candidate pool. Hit@5 rises in parallel, from 86.00% under Hybrid RRF to 94.00%, so the reranker recovers in full the shallow-cutoff penalty that fusion introduces at $K = 5$ and restores parity with BM25 at that depth. Table 2 does separate the two stages, since rows 3 and 4 differ only by the reranker; what it does not separate is the contribution of the dense/RRF arm under reranking, which would require a BM25-plus-Cross-Encoder arm (Section 7). None of the pairwise differences in Table 2 is statistically significant at this benchmark size, and Section 7 shows that none could be; the ordering reported here is therefore descriptive of this run rather than established.

5.2 Generation Quality and Pedagogical Alignment

The multi-agent generation module was evaluated over all 50 benchmark queries (n = 50), covering every cognitive tier of Table 1, using the Faithfulness and Answer Relevancy protocols

of RAGAS [23] together with the adversarial Hallucination Defense Rate.

Table 3. Pedagogical and generation performance metrics. Faithfulness, Answer Relevancy, and average latency were computed on 50 queries; Hallucination Defense Rate on 4 adversarial probes. End-to-end latency sums the sequential retrieval (Table 2) and synthesis stages.

Metric Dimension	Evaluated Score	Target Baseline	Status
Faithfulness Score (Factuality)	85.05%	$\geq 80.0\%$	Achieved
Answer Relevancy Score	87.94%	$\geq 85.0\%$	Achieved
Hallucination Defense Rate (Adversarial)	50.00% (2/4) [95% CI 15.0–85.0]	$\geq 50.0\%$	At target (n = 4; not statistically established)
Average Generation Latency (persona synthesis stage)	1.66 s	—	Sub-two-second synthesis
End-to-End Latency per Query (retrieval + reranking + synthesis)	≈ 3.52 s	≤ 3.00 s	Above target

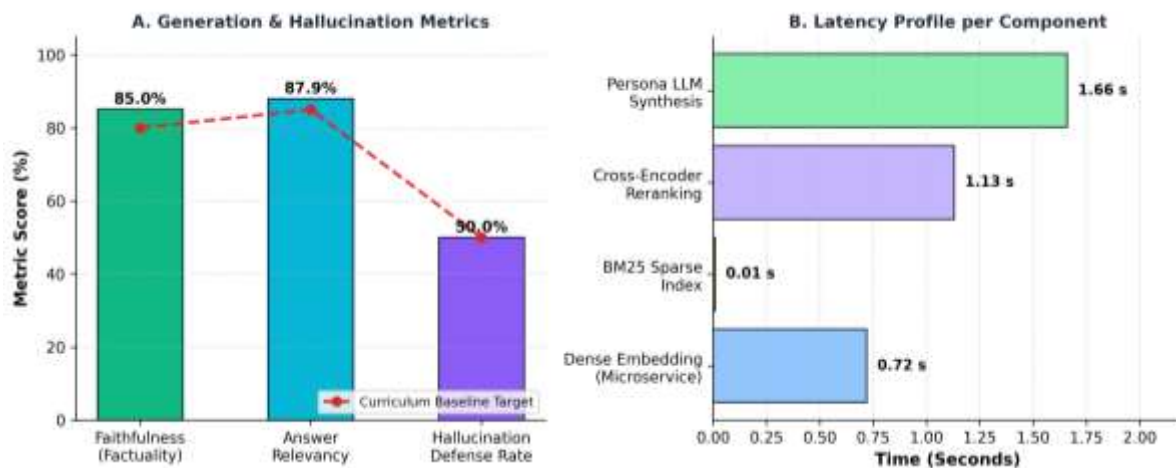


Figure 4. (A) Generation and hallucination metrics against the curriculum baseline target. (B) Per-component latency profile: Persona LLM synthesis (1.66 s) and cross-encoder reranking (1.13 s) dominate the end-to-end budget of approximately 3.52 s, while BM25 lookup is effectively free (0.01 s) and dense embedding accounts for 0.72 s.

Analysis of Generation Performance:

- **High Factual Grounding (85.05%):** The synthesis agent strictly adheres to retrieved contexts. Claims regarding dates, historical figures, and strategic consequences closely match the textbook ground truth, in line with the general finding that RAG substantially curbs hallucination relative to closed-book generation [1], [2].

- **Persona Consistency:** Responses maintain the warm, accessible persona of “Cụ Rùa Thông Thái”, eliminating complex adult vocabulary while reinforcing national pride and historical curiosity an outcome facilitated by the programmable persona and safety rails described in Section 3.4 [5].

- **Robustness & Latency:** Persona synthesis averages 1.66 s per query, but this is the generation stage alone. Adding the retrieval stage of Table 2 gives an end-to-end budget of approximately 3.52 s per query, of which cross-encoder

reranking accounts for 1.13 s and dense embedding for 0.72 s (Figure 4B). The system

is therefore usable for turn-taking classroom interaction but exceeds the 3.00 s target adopted here, and the reranking and dense-embedding stages are the two places where that budget can realistically be reduced.

5.3 Qualitative Case Studies and Error Analysis

Case Study 1: Factual Recall with Entity Ambiguity

Query (Q01): “*Nhà nước đầu tiên trong lịch sử Việt Nam có tên là gì và do ai đứng đầu?*”

Ground Truth: Nước Văn Lang, do Vua Hùng (Hùng Vương) đứng đầu.

Proposed Retrieval: Top-1 chunk retrieved from history_4.md (Unit 1, Lesson 1: Nước Văn Lang).

Generated Answer: “*Chào cháu nhà sử học nhí! Nhà nước đầu tiên trong lịch sử nước ta có tên là Văn Lang, do các Vua Hùng đứng đầu đấy...*”

Evaluation: Hit@1 = True, Faithfulness = 0.94, Relevancy = 0.92.

Case Study 2: Adversarial / Counterfactual Trap

Query (Q47): “*Bác Hồ đã đọc bản Tuyên ngôn Độc lập tại Quảng trường Ba Đình vào năm 1975 có đúng không?*”

Trap: Fuses Ba Dinh Square (1945) with the year of Southern Liberation (1975).

System Defense: The Hybrid Retriever fetched contexts from Lesson 4 (Grade 5): “*Ngày 2-9-1945, Bác Hồ đọc Tuyên ngôn Độc lập khai sinh ra nước Việt Nam Dân chủ Cộng hòa*”.

Agent Output: Correctly refutes the premise, clarifying that the Declaration occurred on September 2, 1945, while 1975 was the year of the Great Spring Victory.

Evaluation: Hallucination Defended = True.

6. Pedagogical Implications and Practical Applications

The integration of our Agentic RAG system into primary education is intended to support three pedagogical functions. These are motivated by, rather than demonstrated by, the present offline evaluation; note that the systematic review drawn on here surveys higher education rather than primary schooling [8]:

- **Inquiry-Based Learning (intended):** Elementary students would be able to explore historical events through natural questions without fear of being judged, promoting intrinsic motivation.
- **Teacher Support Tool (intended):** Educators would be able to leverage the QuizAgent to automatically generate curriculum-aligned review quizzes and discussion prompts tailored to each textbook chapter.
- **Safety Constraints:** Programmable guardrails are designed to reduce hallucinated, inappropriate, or politically biased output. The present evaluation measures a 50.00% Hallucination Defense Rate over four adversarial probes (Table 3) and therefore does not establish a guarantee [5].

7. Limitations and Future Work

While our framework demonstrates strong empirical performance, several avenues for future research remain:

- **Knowledge Graph Integration (GraphRAG):** Cross-dynasty relationships (e.g., comparing naval tactics of Ngo Quyen in 938 versus Tran Hung Dao in 1288) can be further enriched using graph-based retrieval over entity triplets, in the

spirit of query-focused, community-summarized graph retrieval [20].

- **Domain-Specific Embedding Fine-Tuning:** Fine-tuning Vietnamese dense embeddings, building on monolingual pretrained representations such as PhoBERT, on Sino-Vietnamese historical terminology (“Hịch tướng sĩ”, “Thái úy”, “Chiếu dời đô”) will close the gap in dense-only retrieval observed in Section 5.1 [9].

- **Multimodal Extension:** Incorporating historical maps, battle diagrams, and museum artifacts into the retrieval index to support visual question answering for young learners.

- **English-Trained Reranker on Diacritic-Rich Vietnamese:** The cross-encoder used here (ms-marco-MiniLM-L-6-v2) was trained on English MS MARCO passages. Its tokenizer is not adapted to Vietnamese diacritics or Sino-Vietnamese compounds, so part of the residual ranking error in Section 5.1 is likely attributable to this domain and language mismatch; a Vietnamese-trained or multilingual reranker is a necessary comparison.

- **Statistical Power of the Benchmark:** With $N = 50$ retrieval queries and only four adversarial probes, the 95% confidence intervals reported in Tables 2 and 3 are wide and overlap across configurations. The limitation is stronger than overlapping intervals alone, and it can be stated exactly. For a paired binary outcome the exact McNemar test depends only on the discordant counts b and c , where b is the number of queries answered correctly by the baseline but not by the proposed pipeline and c the converse. Table 2 does not report b and c separately, but their difference $c - b$ is fixed by the two totals, so an upper bound on the attainable evidence can be obtained by minimising the exact p -value over all admissible b . Under the most favourable admissible assignment ($b = 0$), the Hit@3 improvement of the proposed pipeline over BM25, 43/50 against 47/50, yields $p = 0.125$; the Hit@1 improvement, 36/50 against 37/50, yields $p = 1.000$; and Hit@5 is identical at 47/50 in both configurations. No pairwise comparison in Table 2 can therefore reach $p < 0.05$ at $N = 50$, whatever the per-query outcomes turn out to be. Holding the observed effect sizes fixed, a paired design would require on the order of 150 queries to establish the Hit@3 difference and on the order of two thousand to establish the Hit@1 difference, so the Hit@3 comparison is the one worth powering and the Hit@1 difference of two percentage points, a single query in fifty, should not be carried forward as a finding. Scaling the benchmark and reporting paired significance tests are required before the observed ordering of architectures can be treated as established; for retrieval comparisons of this kind the randomisation and bootstrap tests are the appropriate choices, the paired t -test being an acceptable approximation at the 50-topic sample size used in TREC evaluations and matched here, and the Wilcoxon signed-rank and sign tests being explicitly discouraged [31].

- **Unsegmented Lexical Indexing:** The BM25 index is built over whitespace-delimited tokens. Because many Vietnamese meaningful words (compound words) span two or more syllables, this fragments multi-syllable historical proper nouns into individually frequent terms. Rebuilding the sparse index

over word-segmented text, and re-running the ablation of Section 5.1 against it, is the most immediate experiment available to the retrieval stage and may change the relative standing of the sparse and fused configurations. The direction of that change should not be assumed, however. Segmentation accuracy is reported on general-domain newswire, whereas Sino-Vietnamese proper nouns and campaign names are precisely the strings on which segmenters disagree, and a query segmented differently from the document that contains the same entity will fail to match under BM25 where unsegmented syllable matching would have succeeded. The segmenter should therefore be validated on the historical terminology of this corpus before the index is rebuilt, and the segmented configuration reported as an additional ablation arm that may lose rather than as an assumed improvement.

• **Ablation Granularity:** Table 2 isolates the incremental effect of reranking over the fused candidate pool, since rows 3 and 4 differ only by the Cross-Encoder, but it does not isolate the contribution of the dense/RRF arm under reranking. A BM25-plus-Cross-Encoder configuration run on the same 50 queries is therefore required, and we leave it to future work; until it is run, the necessity of the dense arm is not established, and the present results should be read as evaluating the pipeline as configured. The same run would supply the per-query cross-tabulation of sparse and dense successes that Section 5.1 currently lacks. A related extension is self-reflective retrieval, in which the generator itself decides when to re-retrieve and whether its draft is grounded [22]; this targets the same failure mode as the Hallucination Defense Rate and would be evaluated on an expanded adversarial set rather than the four probes used here.

• **Expanded Multi-Agent Safety Auditing:** Systematically re-teaming the Router, Quiz, and Roleplay agents against adversarial prompts, following emerging best practices for programmable-rail LLM applications and for multi-agent system evaluation [4], [5].

8. Conclusion

In this work, we presented an Agentic RAG framework specifically engineered for Vietnamese Elementary History Education. By coupling Structural Breadcrumb Semantic Chunking with a two-stage Hybrid Retrieval (BM25 + Dense RRF) and Cross-Encoder Reranking pipeline [13]–[15], the system achieves a Hit Rate@3 of 94.00% and an MRR@5 of 0.8233 across a 50-question golden curriculum benchmark. The persona-driven multi-agent architecture [4], [5] delivers verified historical knowledge with 85.05% Faithfulness, at a synthesis latency of 1.66 s and an end-to-end latency of approximately 3.52 s per query, evaluated under the RAGAS protocol [23]. Within the limits of a single-corpus, 50-question benchmark, this research provides preliminary evidence that carefully architected Agentic RAG systems can mitigate LLM hallucination [1], [2], and points toward a safer and more reliable tool for young history learners, pending classroom validation.

References

[1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, W. Dai, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM*

Comput. Surv., vol. 55, no. 12, Art. no. 248, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

- [2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33. Red Hook, NY, USA: Curran Associates, 2020, pp. 9459–9474.
- [3] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 6769–6781, doi: 10.18653/v1/2020.emnlp-main.550.
- [4] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest, and X. Zhang, "Large language model based multi-agents: A survey of progress and challenges," in *Proc. 33rd Int. Joint Conf. Artificial Intelligence (IJCAI), Survey Track*, 2024, pp. 8048–8057, doi: 10.24963/ijcai.2024/890.
- [5] T. Rebedea, R. Dinu, M. N. Sreedhar, C. Parisien, and J. Cohen, "NeMo Guardrails: A toolkit for controllable and safe LLM applications with programmable rails," in *Proc. 2023 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*, 2023, pp. 431–445, doi: 10.18653/v1/2023.emnlp-demo.40.
- [6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," 2024, arXiv:2312.10997.
- [7] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: A systematic survey," *Comput. Educ.: Artif. Intell.*, vol. 8, Art. no. 100417, 2025, doi: 10.1016/j.caeai.2025.100417.
- [8] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education — where are the educators?," *Int. J. Educ. Technol. Higher Educ.*, vol. 16, Art. no. 39, 2019, doi: 10.1186/s41239-019-0171-0.
- [9] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1037–1042, doi: 10.18653/v1/2020.findings-emnlp.92.
- [10] K. V. Nguyen, D.-V. Nguyen, A. G.-T. Nguyen, and N. L.-T. Nguyen, "A Vietnamese dataset for evaluating machine reading comprehension," in *Proc. 28th Int. Conf. Computational Linguistics (COLING)*, 2020, pp. 2595–2605, doi: 10.18653/v1/2020.coling-main.233.
- [11] N. Reimers and I. Gurevych, "Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language*

- Processing (EMNLP-IJCNLP), 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
- [12] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," in Proc. 35th Conf. Neural Information Processing Systems: Datasets and Benchmarks Track (Round 2), 2021, arXiv:2104.08663.
- [13] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/15000000019.
- [14] G. V. Cormack, C. L. A. Clarke, and S. Büttcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in Proc. 32nd Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2009, pp. 758–759, doi: 10.1145/1571941.1572114.
- [15] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin, "Document ranking with a pretrained sequence-to-sequence model," in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 708–718, doi: 10.18653/v1/2020.findings-emnlp.63.
- [16] R. Nogueira and K. Cho, "Passage re-ranking with BERT," 2019, arXiv:1901.04271.
- [17] O. Khattab and M. Zaharia, "ColBERT: Efficient and effective passage search via contextualized late interaction over BERT," in Proc. 43rd Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2020, pp. 39–48, doi: 10.1145/3397271.3401075.
- [18] T. Formal, B. Piwowarski, and S. Clinchant, "SPLADE: Sparse lexical and expansion model for first stage ranking," in Proc. 44th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2021, pp. 2288–2292, doi: 10.1145/3404835.3463098.
- [19] J. Lin, R. Nogueira, and A. Yates, *Pretrained Transformers for Text Ranking: BERT and Beyond, Synthesis Lectures on Human Language Technologies*. San Rafael, CA, USA: Morgan & Claypool, 2021.
- [20] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, and J. Larson, "From local to global: A Graph RAG approach to query-focused summarization," 2024, arXiv:2404.16130.
- [21] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in Proc. 11th Int. Conf. Learning Representations (ICLR), 2023, arXiv:2210.03629.
- [22] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in Proc. 12th Int. Conf. Learning Representations (ICLR), 2024, arXiv:2310.11511.
- [23] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAS: Automated evaluation of retrieval augmented generation," in Proc. 18th Conf. European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024, pp. 150–158, arXiv:2309.15217.
- [24] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "VnCoreNLP: A Vietnamese natural language processing toolkit," in Proc. 2018 Conf. North American Chapter of the Association for Computational Linguistics: Demonstrations, 2018, pp. 56–60, doi: 10.18653/v1/N18-5012.
- [25] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4512–4525, doi: 10.18653/v1/2020.emnlp-main.365.
- [26] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *Advances in Neural Information Processing Systems*, vol. 33. Red Hook, NY, USA: Curran Associates, 2020, arXiv:2002.10957.
- [27] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang, "MS MARCO: A human generated machine reading comprehension dataset," 2016, arXiv:1611.09268.
- [28] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, F. Xia, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2022, arXiv:2201.11903.
- [29] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl, *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: Cognitive Domain*. New York, NY, USA: Longmans, Green, 1956.
- [30] L. W. Anderson and D. R. Krathwohl, Eds., *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, Complete ed. New York, NY, USA: Longman, 2001.
- [31] M. D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in Proc. 16th ACM Conf. Information and Knowledge Management (CIKM), 2007, pp. 623–632, doi: 10.1145/1321440.1321528.