CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2026.056

Vichat: summary-augmented contextual retrieval for Vietnamese intangible cultural heritage question answering

Tran Minh Tan, Tran Quoc Khang, Nguyen Thai Nghe, and Do Thanh Nghi*

College of Information and Communication Technology, Can Tho University, Viet Nam

**Corresponding author (dtnghi@ctu.edu.vn)*

Article info.

Received 25 Jun 2026
Revised 8 Sep 2026
Accepted 7 Oct 2026

Keywords

Contextual retrieval, intangible cultural heritage, question answering, retrieval-augmented generation, Vietnamese NLP

ABSTRACT

Retrieval-augmented generation over cultural heritage archives is undermined by semantic isolation; fixed-size chunking strips the topic, temporal, and geographic context that makes a passage interpretable, so a chunk reading recognized as national heritage in 2014 no longer identifies which festival or province it describes. Contextual Retrieval repairs this by having an LLM contextualize each chunk, but at one LLM call per chunk. We propose Summary-Augmented Contextual Retrieval (SACR), which summarizes each document once and prepends that single summary to all of its chunks before indexing, reducing cost to one call per document. We further release ViCH-QA, an expert-validated benchmark of 4,719 Vietnamese question-answer (QA) pairs over 159 documents spanning 18 intangible cultural heritage (ICH) topics of Southern Vietnam. Evaluating lexical, dense, and hybrid retrievers under two Vietnamese word segmenters and an unsegmented baseline, we find that SACR raises NDCG@10 in all six matched configurations tested, most sharply for unreranked BM25 (+0.0394), reaching 0.8840 at best. The improvement reflects an inherent trade-off: SACR concentrates relevant passages at the top of the ranking while slightly reducing Hit@10, which is favorable for downstream generation but less suited for broad recall-oriented retrieval. Across six open-source generator backbones, the optimal pipeline is invariably SACR with reranking and few-shot prompting, peaking at a token-F1 of 0.5560; since SACR's cost is paid offline and few-shot prompting is faster than chain-of-thought, our most accurate configuration is also among the cheapest to serve.

1. INTRODUCTION

Preserving *Intangible Cultural Heritage (ICH)* safeguards a society's historic narratives, collective memory, and spiritual identity (Peoples, 2014). Vietnam's living heritage is exceptionally rich, spanning community celebrations such as the Khmer *Ok Om Bok* festival, specialized craftsmanship like the wooden boatbuilding of *Long Định*, and worship traditions such as the *Lễ cúng Việc lề* ceremony of the Southern plains. Their transmission is increasingly threatened by industrialization and demographic shifts, and even where these assets have been digitized the knowledge typically resides in isolated databases, museum archives, or unstructured administrative documents, inaccessible to researchers, tourists, and the general public.

Conversational *Artificial Intelligence (AI)* offers an intuitive interface to such repositories, shifting interaction from keyword matching to natural

language (Brown et al., 2020). Operating on parametric knowledge alone, however, *Large Language Models (LLMs)* hallucinate (Ji et al., 2023) struggling to maintain historical timelines or differentiate localized practices. *Retrieval-Augmented Generation (RAG)* mitigates this by grounding outputs in retrieved passages (Lewis et al., 2020) but inherits a defect from ingestion: fixed-token chunking strips overarching themes, timelines, and geographic markers. Under this *semantic isolation*, a passage reading *"the festival was recognized as a national heritage in 2014"* no longer identifies which festival or province it describes, producing retrieval failures, entity drift, and downstream hallucination (Trivedi et al., 2023). A natural remedy is *Contextual Retrieval*, which asks a lightweight LLM to write a situating preamble for every chunk before indexing; this restores context, but at one LLM invocation per chunk, which becomes the dominant expense for collections that must be re-indexed as they grow.

We therefore present **ViCHat**, a Vietnamese ICH question answering system built around **Summary-Augmented Contextual Retrieval (SACR)**. Rather than contextualizing each chunk independently, SACR summarizes each document once and prepends that single summary to every child chunk prior to indexing, reducing cost from one LLM call per chunk to one per document while still injecting document-wide topic, temporal, and geographic context into each indexed unit. We evaluate SACR against lexical, dense, and hybrid retrievers under three Vietnamese segmentation strategies, with and without cross-encoder reranking, then test whether the gains survive end-to-end across six open-source generator backbones served via vLLM (Kwon et al., 2023). A consistent theme emerges: inexpensive offline context injection paired with simple few-shot prompting outperforms heavier online machinery.

Research Questions: (RQ1) How do Vietnamese word segmentation strategies and retriever families affect retrieval accuracy on cultural heritage queries? (RQ2) Does prepending a single document-level summary to each chunk improve retrieval, and does the gain hold across retriever families and reranking regimes? (RQ3) Do those gains survive end-to-end across generator backbones, and at what latency cost relative to chain-of-thought prompting?

Contributions:

C1: ViCH-QA, an expert-validated benchmark of 4,719 Vietnamese QA pairs over 159 curated documents and 18 Southern ICH topics, spanning single-hop, single-document compositional, and cross-document questions.

C2: SACR, a contextual retrieval scheme requiring one LLM call per document rather than per chunk, which improves NDCG@10 in all six matched retriever configurations we test.

C3: An end-to-end validation across six generator backbones showing the optimal pipeline is invariably SACR with reranking under few-shot prompting, and a latency analysis showing that this most accurate configuration is also among the cheapest to serve, since its cost is paid offline.

Paper organization: The remainder of this paper is organized as follows. Section 2 reviews related work on digital cultural heritage preservation, retrieval-augmented generation, semantic isolation, and prompting strategies. Section 3 describes the construction and characteristics of the ViCH-QA benchmark. Section 4 formalizes the ViCHat retrieval and generation pipeline, including SACR, retriever formulations, reranking, and evaluation settings. Section 5 reports the retrieval and generation results and analyzes the efficiency-quality trade-offs. Finally, Section 6 concludes the paper and discusses limitations and future research directions.

2. RELATED WORK

Digital Preservation of Vietnamese Cultural Heritage: ICH preservation has traditionally relied on manual archiving, ethnographic recording, and regional museum catalogues (Peoples, 2014); in Southern Vietnam, performative arts such as *Đờn ca tài tử Nam Bộ*, ancestral rituals like *Lễ cúng Việc lề*, and traditional crafts face severe transmission challenges. Early computational work digitized unstructured documents and built tourism-centric QA systems on semantic networks and early neural architectures to extract destinations and historical timelines (Do et al., 2021).More recently, multimodal frameworks such as VietMEAgent (Nguyen et al., 2025) combine object detection with structured program generation for culturally specific queries. These systems nonetheless rest on hand-crafted knowledge graphs or closed-domain databases, and are fragile when confronted with open-ended queries that link historical and ritual context across documents.

RAG, Semantic Isolation, and Prompting Paradigms: RAG is the de facto standard for knowledge-intensive NLP, grounding LLM outputs in retrieved evidence to curb hallucination (Lewis et

al., 2020; Ji et al., 2023). Retrievers divide into sparse lexical matchers such as BM25 (Robertson & Zaragoza, 2009) and dense bi-encoders, with hybrid fusion via *Reciprocal Rank Fusion (RRF)* (Cormack et al., 2009) mitigating vocabulary mismatch; all of them inherit whatever the ingestion stage produces. Because standard ingestion splits documents into fixed-size chunks, it severs the metadata, regional context, and historical framing that make a passage interpretable, a failure mode termed *semantic isolation* (Tao et al., 2025). *Contextual Retrieval* has a lightweight LLM write a situating preamble per chunk before indexing, at one LLM call per chunk; SACR retains the document-level grounding while collapsing this to one call per document. To our knowledge, neither variant has been systematically evaluated for Vietnamese cultural heritage. Both belong to a family of remedies that intervene at indexing time, ensuring each retrievable unit already carries enough context that the facts a compositional question needs co-occur once retrieved. A second family instead intervenes at generation time via *Chain-of-Thought (CoT)* (Wei et al., 2022) or few-shot exemplars (Brown et al., 2020); iterative retrieve-and-reason loops push this furthest but suffer entity drift, where an early retrieval error propagates through subsequent queries (Sun et al., 2026), while multiplying inference cost. The two families are rarely measured against each other, and rarely under resource constraints. We hold generation simple, invest at indexing time, and quantify how far that choice carries.

3. THE ViCH-QA DATASET

We construct ViCH-QA from a manually curated corpus of 159 articles spanning 18 ICH topics of *Southern Vietnam*, covering festivals (e.g. *Ok Om Bok*), performing arts (e.g. *Đờn ca tài tử*), crafts (e.g. *Nghề đóng xuồng ghe Long Định*), and ancestral rites.

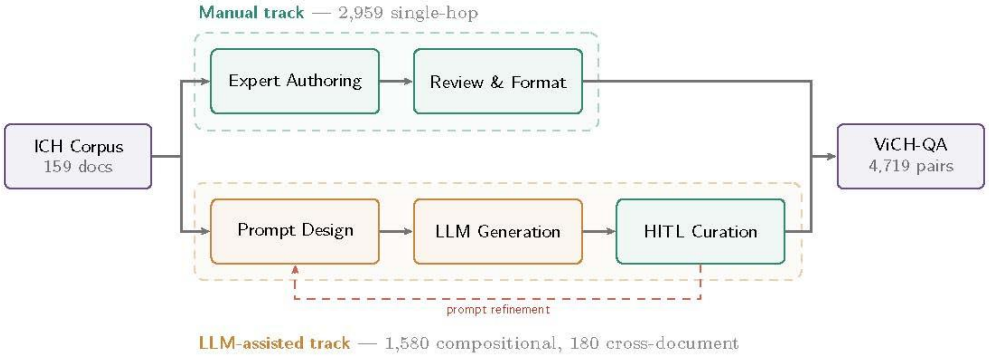


Figure 1. ViCH-QA construction. A manual track produces single-hop pairs; an LLM-assisted track with human-in-the-loop curation produces compositional and cross-document pairs

As shown in, construction follows two tracks, both ending in human-in-the-loop review, and yields three question types of increasing difficulty. The first is fully manual and produces the 2,959 single-hop pairs (62.7%), whose answers lie within a single chunk and which serve as the baseline for standard retrieval. The second is LLM-assisted and produces the two harder categories with DeepSeek-V4-Flash: 1,580 single-document compositional questions (33.5%), requiring synthesis over several chunks of one parent document, and 180 cross-document questions (3.8%), requiring evidence from distinct source documents. Both generated categories are then thoroughly reviewed and edited by domain experts, who discard hallucinated content, repair unnatural phrasing, and verify that each stated fact is strictly grounded in the cited texts, mitigating potential generator bias from DeepSeek-V4-Flash. Across the 4,719 questions, there are 4,940 source citations to parent documents, averaging 1.05 documents per question, up to four. Specifically, each single-hop question and single-document compositional question cites exactly one parent document, the latter requiring synthesis across multiple internal chunks of that document, while the 180 cross-document questions cite between 2 and 4 distinct documents (401 citations total, averaging 2.23 documents per question); Figure 2 summarizes the topical and structural composition. One cross-document question, for instance, asks which two festivals join *Ok Om Bok* among the three largest Khmer festivals and in which year it was recognised as national intangible heritage in Trà Vinh, whose reference answer (*Chôl Chnăm Thmây and Sen Đôlta; 2014*) spans two sources, neither sufficient alone.

Table 1. ViCH-QA length statistics over all 4,719 question-answer pairs, under raw, PyVi and Underthesea (UTS) tokenisation

	Question				Answer			
	Chars	Raw	PyVi	UTS	Chars	Raw	PyVi	UTS
Mean	119.9	27.40	21.24	20.44	208.4	46.71	39.52	38.00
Std.	29.3	6.50	5.13	5.03	85.6	19.01	17.08	16.42
Median	119	27	21	20	202	45	38	36
Max	276	63	56	55	756	171	158	154

Lexical effect of word segmentation. Table 1 quantifies a property of Vietnamese central to the Results section: segmentation compresses the token stream by roughly 25%, from 27.40 raw whitespace syllables per question to 20.44 under Underthesea, by merging syllables into multi-syllabic lexical units. The units it forms are the ones that matter here: the most frequent compounds in the segmented corpus are *lễ hội* (festival), *văn hóa* (culture), and *di sản* (heritage), precisely the discriminative terms for this domain, whereas raw tokenization fragments them into common, low-information syllables.

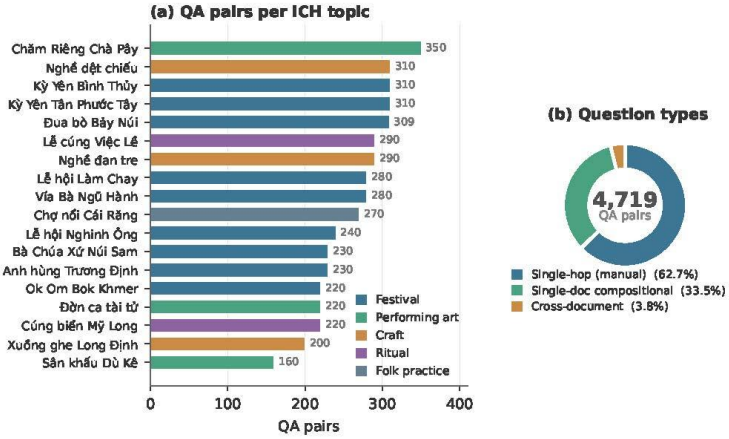


Figure 2. ViCH-QA composition

(Notes: (a) QA pairs per ICH topic, colored by heritage category (b) distribution over the three question types)

4. METHODOLOGY

Our system (Figure 3) decomposes into an offline indexing stage and an online retrieval-generation stage, which we formalize in the remainder of this section.

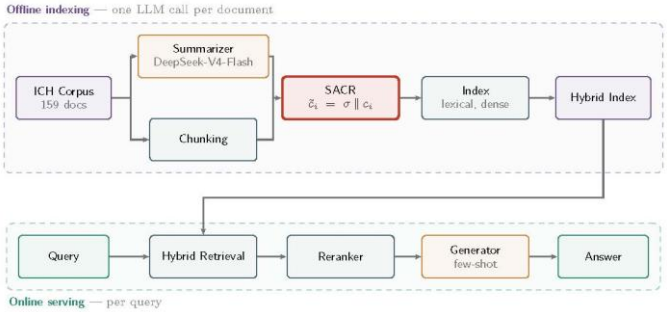


Figure 3. Overview of the ViCHat RAG pipeline

(Notes: Top: offline indexing, where each document is chunked and a single DeepSeek-V4-Flash-generated document summary is prepended to every child chunk (SACR) before embedding and BM25 indexing Bottom: online retrieval and generation, where hybrid retrieval, cross-encoder reranking, and a few-shot prompted generator produce the final grounded answer.)

4.1. Retriever Formulations

Let $\mathcal{D} = \{d_1, \dots, d_N\}$ be the corpus of N documents. Each d_j is split into M_j contiguous chunks $\mathcal{C}_j = \{c_{j,1}, \dots, c_{j,M_j}\}$, and $\mathcal{C} = \bigcup_j \mathcal{C}_j$ is the full chunk pool. Given a query q, a retriever ranks every $c \in \mathcal{C}$ by a relevance score $S(q, c)$, and we write $\mathrm{Ret}(q, \mathcal{C})$ for the top-k chunks it returns.

Lexical retrievers: We evaluate BoW and TF-IDF as baselines and Okapi BM25 (Robertson & Zaragoza, 2009) as the main lexical model, which normalizes for document length and saturates term frequency:

$$S_{\mathrm{BM25}}(q, c) = \sum_{w \in q \cap c} \mathrm{IDF}(w) \cdot \frac{f_{w,c}\,(k_1 + 1)}{f_{w,c} + k_1\left(1 - b + b\,\frac{|c|}{\mathrm{avgdl}}\right)} \quad (1)$$

where $f_{w,c}$ is the frequency of term w in chunk c, $|c|$ its length in tokens, $avgdl$ the mean chunk length over $\mathcal{C}$, and $k_1 = 1.5$, $b = 0.75$ control term saturation and length normalisation. The inverse document frequency is $\mathrm{IDF}(w) = \log\big(1 + (|\mathcal{C}| - n_w + 0.5)/(n_w + 0.5)\big)$ with n_w the number of chunks containing w; both $|c|$ and n_w are computed on the indexed text, which under SACR means the summary-augmented chunk.

Dense and hybrid retrievers: Let $\mathrm{Enc}(\cdot)$ map a text to a d-dimensional embedding. Dense retrieval scores by cosine similarity, $S_{\mathrm{dense}}(q, c) = \cos\big(\mathrm{Enc}(q), \mathrm{Enc}(c)\big)$. We compare four encoders, identified throughout by their HuggingFace checkpoints: All-MiniLM-L6-v2 (Dense_{MiniLM}), Vietnamese-SBERT ($\mathrm{Dense}_{viSBERT}$), Embedding-Gemma-300M (Dense_{Gemma}), and Llama-Nemotron-Embed-1B ($\mathrm{Dense}_{Nemotron}$). Hybrid retrieval fuses the two rankings via RRF. Writing $\mathrm{rk}_{\mathrm{lex}}(q, c)$ and $\mathrm{rk}_{\mathrm{dense}}(q, c)$ for the 1-based rank position of c under BM25 and under the dense retriever respectively:

$$S_{\mathrm{hybrid}}(q, c) = \frac{w_{\mathrm{BM25}}}{\mathrm{rk}_{\mathrm{lex}}(q, c) + k_{\mathrm{rrf}}} + \frac{w_{\mathrm{dense}}}{\mathrm{rk}_{\mathrm{dense}}(q, c) + k_{\mathrm{rrf}}} \quad (2)$$

with smoothing constant $k_{rrf} = 60$ and fusion weights $w_{dense} = 1.0$, $w_{BM25} = 0.5$ throughout.

Cross-encoder reranking. A reranker jointly encodes the concatenated pair $[q; c]$ with full cross-attention. Writing $h([q; c])$ for the resulting pooled representation, $\mathbf{w}$ for a learned scoring vector and sig for the logistic function, the reranking score is $S_{\mathrm{CE}}(q, c) = \mathrm{sig}\big(\mathbf{w}^{\top} h([q; c])\big)$, and the candidate pool is re-sorted by it. We use Llama-Nemotron-Rerank-1B.

Definition 1 (Summary-Augmented Contextual Retrieval): For each document d_j, a lightweight LLM produces a single document-level summary *once, offline*: $\sigma_j = \mathrm{Summ}\big(d_j\big)$. Every child chunk is rewritten by concatenation ($\oplus$) with that shared summary, $\tilde{c}_{j,i}^{\,\mathrm{prepend}} = \sigma_j \oplus c_{j,i}$ or $\tilde{c}_{j,i}^{\,\mathrm{append}} = c_{j,i} \oplus \sigma_j$, and indexing and scoring operate on $\tilde{c}_{j,i}$ rather than $c_{j,i}$, so $S_{\mathrm{SACR}}\big(q, c_{j,i}\big) = S\big(q, \tilde{c}_{j,i}\big)$ for any lexical, dense, or hybrid S. Because σ_j is reused across all M_j chunks of d_j, SACR requires $O(|\mathcal{D}|)$ LLM calls rather than the $O(|\mathcal{C}|)$ of per-chunk *Contextual Retrieval*, a reduction by the mean chunks-per-document factor.

4.2. Generation Pipeline Formulations

Given retrieved context $\mathcal{C}_{\mathrm{ret}} = \mathrm{Ret}(q, \mathcal{C})$, we evaluate two prompting paradigms, each in a zero-shot and a few-shot regime, giving four configurations. Let $\mathbf{p}_{\mathrm{van}}$ and $\mathbf{p}_{\mathrm{CoT}}$ be the two prompt templates, each formatting a query and its retrieved context into the model input. Vanilla RAG (Lewis et al., 2020) produces the answer a in a single step, whereas CoT first emits a rationale $Z = [z_1, \dots, z_L]$ of L intermediate steps:

$$a = \mathrm{LLM}\big(\mathbf{p}_{\mathrm{van}}(q, \mathcal{C}_{\mathrm{ret}})\big), \qquad [Z; a] = \mathrm{LLM}\big(\mathbf{p}_{\mathrm{CoT}}(q, \mathcal{C}_{\mathrm{ret}})\big), \quad (3)$$

where $\mathrm{LLM}(\cdot)$ denotes greedy decoding from the generator and $[\,\cdot\,;\cdot\,]$ concatenation (Wei et al., 2022). In the few-shot regime, both are additionally conditioned on a fixed set $\mathcal{E}$ of $n = 2$ in-context demonstrations, held constant across all queries solely as structural format exemplars and excluded from evaluation. Because our system conducts training-free inference on pre-trained backbones without weight updates, no training/testing data leakage occurs. Because $\mathcal{E}$ does not vary with the query, few-shot prompting adds only a constant prompt prefix: it requires no extra retrieval and no additional generation round-trips, which is why it can improve quality while *reducing* latency. We restrict generation to these lightweight paradigms by design, on the hypothesis that context is better supplied once at indexing time via SACR than reconstructed per query through iterative reasoning.

4.3. Evaluation Settings

Setup. All experiments run on a single machine (Intel i7-14700K, 128 GB RAM, NVIDIA RTX 5070 Ti 16 GB). We compare three tokenisations: *raw whitespace* (raw), *PyVi* (PV, rule- and statistics-based), and *Underthesea* (UTS, sequence labelling). SACR summaries are produced offline with DeepSeek-V4-Flash, the same model used for QA generation but for a distinct purpose. Generation is evaluated across six open-source backbones served via vLLM (Kwon et al., 2023): Qwen3.5-4B, Qwen3-4B, Gemma-4-E2B, Gemma-4-E4B, Llama-3.1-8B, and Qwen3-8B, all under post-training quantization (NVFP4 for the Qwen3.5-4B, Qwen3-8B and Llama checkpoints, FP8 for Qwen3-4B, w4a16 QAT for the Gemma-4 pair), spanning three model families and two parameter scales to test whether SACR's benefit is backbone-dependent. All results are computed on the complete set of 4,719 questions with no subsampling; retrieval is averaged over three runs under seeds {42, 100, 2026}, while generation is decoded greedily and reported from a single pass, so small differences between generation configurations should be read with corresponding caution.

Retrieval metrics and relevance definition: Let $\mathcal{R}_k = [r_1, \dots, r_k]$ be the ranked list of top-k retrieved chunks, each originating from parent document $\mathrm{doc}(r_i)$, and let $\mathcal{G}_q \subset \mathcal{D}$ be the gold supporting documents for query q. A chunk r_i (original or summary-augmented) is deemed relevant if and only if its parent document belongs to the ground-truth set ($\mathrm{doc}(r_i) \in \mathcal{G}_q$). Retrieved documents are deduplicated by first occurrence to prevent multiple chunks from the same document from artificially inflating metrics. We report Hit@$k = \mathbb{I}\left[\exists r_i \in \mathcal{R}_k : \mathrm{doc}(r_i) \in \mathcal{G}_q\right]$; MRR@10, the reciprocal rank of the first relevant document; and NDCG@10 under binary relevance, where DCG@$k = \sum_{i \le k} \mathbb{I}\left[\mathrm{doc}(r_i) \in \mathcal{G}_q\right] / \log_2(i+1)$ and IDCG@$k = \sum_{i \le \min(|\mathcal{G}_q|, k)} 1 / \log_2(i+1)$. For multi-evidence evaluation on compositional and cross-document queries, we further compute Complete-Hit@$k = \mathbb{I}\left[\mathcal{G}_q \subseteq \{\mathrm{doc}(r_i)\}_{i \le k}\right]$, successful only when *all* required supporting documents are retrieved, and All-Evidence Recall@$k = \left|\mathcal{G}_q \cap \{\mathrm{doc}(r_i)\}_{i \le k}\right| / \left|\mathcal{G}_q\right|$.

Generation metrics: Let A_q and T_q be the generated and reference answers for query q, and $V(x)$ the multiset of tokens obtained from x by lowercasing, applying Underthesea segmentation, and stripping punctuation. All lexical metrics operate on this segmented stream, so a compound such as *lễ_hội* counts as a single token; scoring raw syllables would inflate n-gram overlap between unrelated texts. We report Token-F1, the harmonic mean of precision and recall under $V(A_q) \cap V(T_q)$; BLEU-4 (Papineni et al., 2002) with uniform n-gram weights and NLTK method-1 smoothing; and ROUGE-L (Lin, 2004), the LCS-based F1 with stemming disabled, since Vietnamese is isolating and an English stemmer would corrupt tokens. Because these three penalise valid paraphrasing and reordering, we add BERTScore (Zhang et al., 2020), which aligns the two answers greedily in embedding space using the scoring encoder's own subword tokenisation rather than $V(\cdot)$. Writing $\mathcal{A}$ and $\mathcal{T}$ for the resulting token sequences of A_q and T_q, and $\hat{\mathbf{v}}_u = \mathbf{v}_u / \|\mathbf{v}_u\|_2$ for the ℓ_2-normalised contextual embedding of token u under encoder Φ, precision and recall are:

$$P_\Phi = \frac{1}{|\mathcal{A}|} \sum_{u \in \mathcal{A}} \max_{t \in \mathcal{T}} \hat{\mathbf{v}}_u^\top \hat{\mathbf{v}}_t, \qquad R_\Phi = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \max_{u \in \mathcal{A}} \hat{\mathbf{v}}_u^\top \hat{\mathbf{v}}_t, \tag{4}$$

and $\mathrm{BS}_\Phi = 2P_\Phi R_\Phi / (P_\Phi + R_\Phi)$ is their harmonic mean. We instantiate Φ with two encoders from different families, EmbeddingGemma-300M (BS_{Gemma}) and Llama-Nemotron-Embed-1B ($\mathrm{BS}_{Nemotron}$), so that reported rankings are not an artefact of a single embedding space.

5. RESULTS

5.1. Retrieval Performance

RQ1: Word segmentation is cheap and reliably aids term-weighted retrieval: *Result:* Block (a) of Table 2 shows lexical retrieval improving monotonically with segmentation quality: TF-IDF rises in NDCG@10 from 0.6708 (raw) to 0.7076 (PV) and 0.7228 (UTS), while BM25 increases from 0.8028 to 0.8275 and 0.8325. *Explanation:* Word segmentation groups individual syllables into discriminative multi-syllabic compounds (Table 1), suppressing spurious unigram matches and concentrating scoring mass on domain-specific cultural entities at negligible runtime cost (BM25 latency increases from 0.85 to only 1.23 ms/query).

Table 2. Retrieval results, averaged over three runs on all 4,719 questions raw/PV/UTS denote raw, PyVi and Underthesea tokenization

Configuration	Hit@5	Hit@10	MRR@10	NDCG@10	ms/q
(a) Lexical baselines: *effect of word segmentation*					
BoW_{raw}	0.5806	0.7112	0.3717	0.4489	0.05
BoW_{PV}	0.6111	0.7402	0.3936	0.4727	0.22
BoW_{UTS}	0.5702	0.6987	0.3737	0.4476	0.64
TF-IDF_{raw}	0.8252	0.9335	0.5923	0.6708	0.07
TF-IDF_{PV}	0.8546	0.9494	0.6342	0.7076	0.24
TF-IDF_{UTS}	0.8712	0.9525	0.6540	0.7228	0.67
BM25_{raw}	0.9241	0.9705	0.7528	0.8028	0.85
BM25_{PV}	0.9371	0.9703	0.7850	0.8275	0.81
BM25_{UTS}	0.9385	0.9695	0.7917	**0.8325**	1.23
(b) Dense bi-encoders					
Dense_{MiniLM}	0.3999	0.5234	0.2550	0.3143	0.61
$\text{Dense}_{viSBERT}$	0.6217	0.7843	0.3996	0.4874	1.64
Dense_{Gemma}	0.8814	0.9570	0.6560	0.7257	2.54
$\text{Dense}_{Nemotron}$	0.9194	0.9645	0.7247	**0.7801**	82.99
(c) Effect of SACR, *no reranking (matched pairs)*					
BM25_{UTS}	0.9385	0.9695	0.7917	0.8325	1.23
with SACR *(prepend)*	0.9566	0.9650	0.8466	**0.8719**	133.38
with SACR *(append)*	0.9557	0.9646	0.8428	0.8687	134.55
$\text{Hybrid}_{Gemma,UTS}$	0.9289	0.9804	0.7266	0.7856	3.81
with SACR *(prepend)*	0.9371	0.9617	0.7456	0.7957	5.00
$\text{Hybrid}_{Nemotron,UTS}$	0.9534	0.9787	0.7696	0.8178	92.20
with SACR *(prepend)*	0.9589	0.9736	0.7783	0.8232	81.97
(d) With cross-encoder *reranking*					
Dense_{Gemma}, rerank	0.9519	0.9761	0.8289	0.8623	199.55
$\text{Dense}_{Nemotron}$, rerank	0.9625	0.9819	0.8321	0.8660	230.20
BM25_{UTS}, rerank	0.9597	0.9797	0.8391	0.8709	205.48
with SACR *(prepend)*	0.9640	0.9741	0.8504	0.8777	159.68
$\text{Hybrid}_{Gemma,UTS}$, rerank	0.9639	0.9880	0.8386	0.8729	205.59
with SACR *(prepend)*	0.9683	0.9793	0.8511	0.8801	164.41
$\text{Hybrid}_{Nemotron,UTS}$, rerank	0.9676	0.9879	0.8368	0.8709	223.45
with SACR *(prepend)*	**0.9739**	0.9833	**0.8556**	**0.8840**	237.60

(Notes: *Rows tinted blue apply SACR; the shaded row is the best configuration overall. Latency is milliseconds per query)*

BoW is the lone exception where UTS (0.4476) trails both PV (0.4727) and unsegmented text (0.4489); lacking IDF weighting, merging syllables into rarer compounds reduces raw term overlap. Among dense encoders (block b), scale and multilingual pre-training dominate: $\text{Dense}_{\text{Nemotron}}$ reaches 0.7801 and $\text{Dense}_{\text{Gemma}}$ 0.7257, compared to 0.4874 for $\text{Dense}_{\text{viSBERT}}$ and 0.3143 for $\text{Dense}_{\text{MiniLM}}$. *Implication:* Exact lexical matching under proper Vietnamese segmentation remains the strongest first-pass signal in this entity-dense archive: segmented BM25 outperforms every non-reranked dense encoder while executing in under 2ms.

RQ2: SACR trades candidate breadth for top-ranked precision: *Result:* Blocks (c) and (d) demonstrate that across all six matched configurations, SACR consistently raises NDCG@10, MRR@10, and Hit@5 (Figure 4a), while moderately lowering Hit@10 (mean −0.0079,

range -0.0045 to -0.0187). Paired significance testing across all 4,719 queries confirms that SACR's NDCG@10 improvements are statistically significant ($p < 0.001$ under both paired t-tests and Wilcoxon signed-rank tests; e.g., $t = 3.521$, $p = 4.34 \times 10^{-4}$ for BM25_{UTS}). Evaluating by question type further clarifies this behavior: (1) On single-hop questions ($N = 2{,}959$), SACR improves NDCG@10 from 0.8576 to 0.8649 with stable Hit@10 (0.9807 vs 0.9784); (2) On single-document compositional questions ($N = 1{,}580$), SACR lifts NDCG@10 from 0.7774 to 0.7884 and Hit@10 from 0.9310 to 0.9335; (3) On cross-document questions ($N = 180$), SACR raises NDCG@10 from 0.6371 to 0.6461 with identical Hit@10 (0.9611). For multi-evidence metrics on cross-document queries, Complete-Hit@10 is 0.6056 for baseline BM25 vs 0.5778 for SACR, and All-Evidence Recall@10 is 0.7921 vs 0.7819. *Explanation:* Prepending a shared document summary makes every chunk self-describing, enabling the retriever to rank the primary passage near rank 1 (MRR@10 on BM25 rises from 0.7917 to 0.8466). However, because child chunks of the same document share the summary prefix, sibling chunks become more similar to one another and compete for the remaining top-10 slots. For cross-document queries, this slightly crowds out evidence from secondary documents in the lower top-10 ranks. *Implication:* For downstream RAG generation, which consumes only the top-k ($k = 3$) passages, concentrating relevance mass at the very top of the ranking is decisively advantageous, whereas recall-oriented tasks require candidate diversification. Compared to Anthropic's per-chunk contextual retrieval requiring $O(|\mathcal{C}|)$ LLM invocations, SACR requires only $O(|\mathcal{D}|)$ calls (saving 68% to 96% of indexing calls depending on chunk density) while providing holistic narrative framing that fixed metadata or title prepending cannot match. The gain is largest for bare BM25 without reranking (+0.0394 NDCG@10), where the prepended summary directly compensates for the lack of a neural reranker. The interaction with reranking is retriever-dependent: the largest reranked gain (+0.0131, $\text{Hybrid}_{\text{Nemotron}}$) exceeds two unreranked gains. Prepending outperforms appending in five of six pairs, and we adopt prepending throughout.

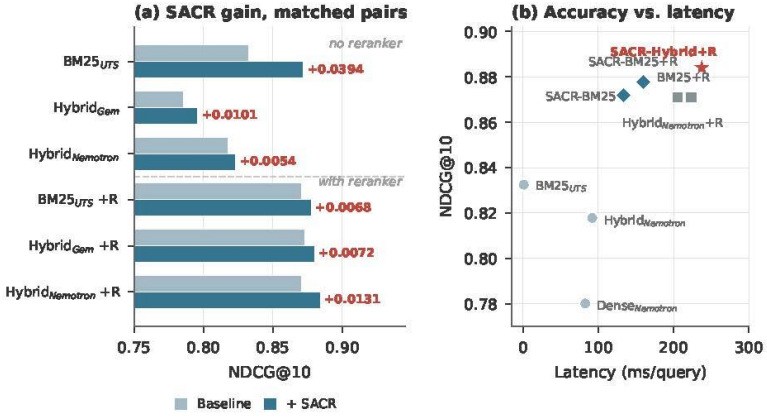


Figure 4. Impact of SACR on Retrieval Performance and Latency

*(**Notes:** (a) SACR raises NDCG@10 across all six matched configurations, showing the largest gain for unreranked BM25. (b) On the accuracy-latency plane, SACR over BM25 (diamonds) dominates reranked baselines (squares) by achieving higher NDCG@10 at lower latency; the best overall configuration (star) delivers the highest accuracy but carries the highest latency.)*

Operating points and efficiency profiling: The most accurate configuration overall is SACR over hybrid retrieval with reranking, reaching Hit@5 = 0.9739, MRR@10 = 0.8556, and NDCG@10 = 0.8840 (237.60ms). A highly efficient alternative, SACR over BM25 with reranking, reaches 0.8777 at 159.68ms, recovering 99.3% of the top score with 33% lower latency (Figure 4b). In contrast, the top configuration yields a marginal gain of 0.0063 NDCG@10 for 78ms of additional latency.

5.2. Generation Performance

Blocks (a)-(b) of Table 3 report the best configuration for each of the six generator backbones; block (c) isolates the effect of prompting on the strongest backbone.

Table 3. Generator Backbones and Prompting Ablation

Generator	Prompting	F1	BLEU	$ROUGE_L$	BS_{Gemma}	$BS_{Nemotron}$	s/q
***(a) 4B-class** backbones, best configuration*							
Qwen3.5-4B	Few-shot vanilla	0.5473	0.2598	0.4644	0.8417	0.7948	7.45
Qwen3-4B	Few-shot vanilla	0.5417	0.2682	0.4651	0.8478	0.8052	10.78
Gemma-4-E2B	Few-shot CoT	0.5522	0.2692	0.4768	0.8414	0.7946	16.55
Gemma-4-E4B	Few-shot vanilla	0.5514	0.2664	0.4715	0.8422	0.7976	12.36
***(b) 8B-class** backbones, best configuration*							
Llama-3.1-8B	Few-shot vanilla	0.4802	0.2217	0.4035	0.8150	0.7645	11.46
Qwen3-8B	Few-shot vanilla	**0.5560**	0.2770	0.4772	**0.8524**	**0.8071**	8.15
***(c) Prompting ablation**, Qwen3-8B, retriever fixed*							
Qwen3-8B	Zero-shot van.	0.5050	0.2404	0.4308	0.8455	0.8057	10.77
Qwen3-8B	Zero-shot CoT	0.5353	0.2598	0.4581	0.8473	0.8018	15.22
Qwen3-8B	Few-shot CoT	0.5552	**0.2779**	**0.4783**	0.8468	0.7981	13.09

*(**Notes:** Blocks (a)-(b) display the best configuration for each generator backbone, where every backbone selects SACR with reranking as its retriever along with a few-shot prompt. Block (c) shows the prompting ablation on Qwen3-8B while holding the retriever fixed. Latency is reported in seconds per query).*

RQ3: SACR's downstream benefit is backbone-independent: *Result:* For all six generator backbones across three distinct model families, two parameter scales, and three quantization regimes, the top-performing pipeline consistently couples SACR with cross-encoder reranking under few-shot prompting; no backbone favored a non-contextualized retriever (Table 3). Downstream F1 ranges from 0.4802 (Llama-3.1-8B) to 0.5560 (Qwen3-8B).} *Explanation:* Supplying SACR-augmented passages ensures that key historical, temporal, and geographic anchors enter the generator's context window, preventing entity hallucination. While Qwen3.5-4B gains +0.0114 F1 (0.4928 to 0.5042) and Gemma-4-E2B lifts F1 from 0.4660 to 0.4741, Qwen3-4B's lexical F1 remains flat (0.4966 vs 0.4961) despite clear improvements in semantic similarity (BS_{Gemma} rises from 0.8357 to 0.8401; $BS_{Nemotron}$ from 0.7943 to 0.8004). *Implication:* Lexical metrics under-report valid stylistic paraphrases; contextual retrieval consistently supplies superior factual grounding across diverse LLM backbones.

Few-shot prompting dominates CoT: *Result:* Adding two fixed in-context exemplars to vanilla prompting yields the largest generation gain: Qwen3-8B F1 surges from 0.5050 to 0.5560, while serving latency *drops* by 24% from 10.77s to 8.15s. In contrast, CoT provides a smaller zero-shot gain (0.5050 to 0.5353) at a 41% latency premium, and yields no benefit over few-shot vanilla prompting (0.5552 vs 0.5560) while increasing latency by 60%. *Explanation:* The fixed demonstrations anchor concise, structured formatting, suppressing verbose preambles and repetitive hedging, which saves far more generation tokens than the prefix adds. Only the smallest model (Gemma-4-E2B) benefits from few-shot CoT, where explicit reasoning compensates for lower capacity. *Implication:* Offline context enrichment via SACR combined with lightweight few-shot prompting establishes a highly performant and serving-efficient RAG deployment.

What semantic isolation looks like. The failure mode motivating SACR is visible in a representative error from a non-contextualised retriever.

Example 1: Retrieval failure caused by semantic isolation

Question: Lễ hội Ok Om Bok được vinh danh là di sản văn hóa phi vật thể quốc gia vào những năm nào và ở những tỉnh nào? (*In which years and provinces was Ok Om Bok recognised as national intangible heritage?*)

Reference: Trà Vinh, 2014 · Sóc Trăng, 2015

System output: Trà Vinh, 2014 · Lễ hội Cúng biển Mỹ Long, 2013 (*Sóc Trăng missing; the second date belongs to a different festival*)

The Sóc Trăng passage is in the corpus, but as an isolated chunk it reads roughly as "*recognised as national intangible heritage in 2015*" without naming the festival, leaving it indistinguishable from many similar recognition statements; it never enters the candidate pool, and the generator, handed one on-topic and one off-topic date, fluently merges them. No reasoning over the retrieved set can recover a passage that was never retrieved, so generation-time interventions cannot help here. Ensuring the passage carries its parent document's subject at indexing time is what resolves it, which is

SACR's mechanism, and explains why its largest gains fall on MRR@10.

Why the lexical scores look low: The absolute BLEU and ROUGE-L values in Table 3 understate answer quality, as the example below shows: asked to name at least three art forms of *Ok Om Bok*, our system returns all four present in the reference, but as a numbered list, reordered, with brief glosses and one extra correct detail.

Example 2: A fully correct answer that the lexical metrics penalise
Question: Hãy nêu ít nhất ba loại hình nghệ thuật được trình diễn trong lễ hội Ok Om Bok. (*Name at least three art forms performed at Ok Om Bok.*)
Reference: múa trống Chhay-dăm, múa rô băm, nhạc ngũ âm, múa Răm Vông
System output: 1. múa Răm Vông (điệu múa vòng tròn tập thể); 2. dàn nhạc ngũ âm; 3. múa trống Chhay-dăm (múa trống truyền thống); 4. múa rô băm; 5. đua ghe ngo - *all four reference items recovered, plus one further correct detail (item 5)*
Why it scores poorly: the parenthetical glosses dilute BLEU's *n*-gram precision; the numbered list and the reordering break the longest common subsequence ROUGE-L measures; and the surface variant "dàn nhạc ngũ âm" for "nhạc ngũ âm" counts as a miss.

Hence we report BERTScore under two encoders alongside the lexical metrics and rest the argument on cross-system rankings rather than absolute lexical values.

Execution efficiency: Retrieval latency is negligible besides generation: bare lexical retrievers run under 1.3ms/query, neural components raise this to 83-92ms and to 160-238ms once reranking is added, while generation takes 7-17s/query. The generator, not the retriever, governs end-to-end responsiveness, so the 78ms separating our cheapest and most accurate SACR operating points is under 1% of a single generation call and that choice should be made on accuracy alone. The practically important result is that the most accurate pipeline is also among the fastest: few-shot vanilla prompting on Qwen3-8B gives the study's best F1 (0.5560) at 8.15s, against 13.09-15.22s for CoT variants of equal or lower quality. Combined with SACR's summarization running once per document offline, the recommended configuration carries no inference-time penalty for its accuracy.

6. CONCLUSION

We presented **ViCHat**, a Vietnamese ICH question answering system built around **Summary-Augmented Contextual Retrieval (SACR)**, which counteracts semantic isolation by summarizing each source document once and prepending that summary to every child chunk before indexing, and we release ViCH-QA, an expert-validated benchmark of 4,719 QA pairs over 18 Southern Vietnamese ICH topics. Three findings emerge. First, Vietnamese word segmentation is a near-free and reliable gain for term-weighted lexical retrieval, lifting BM25 NDCG@10 from 0.8028 to 0.8325 for well under a millisecond per query, though it does not help retrievers lacking term weighting. Second, SACR raises NDCG@10, MRR@10 and Hit@5 in all six matched configurations tested, with the largest gain (+0.0394) on unreranked BM25 and the best configuration reaching 0.8840, by concentrating relevant passages at the top of the ranking at a small cost in Hit@10. Third, the benefit survives end-to-end: across six generator backbones spanning three families, the optimal pipeline always paired SACR and reranking with few-shot prompting, peaking at F1 of 0.5560 on Qwen3-8B. Because SACR's cost is paid once offline per document and few-shot prompting is faster than CoT, the recommended configuration is also among the least expensive to serve.

Limitations and future work: The corpus comprises 159 documents and 494 chunks, where SACR's one-call-per-document economy reduces indexing API calls by 68% compared to per-chunk contextualization ($O(|\mathcal{D}|)$ vs $O(|\mathcal{C}|)$); while our theoretical complexity analysis establishes this cost advantage, a direct head-to-head empirical quality comparison against per-chunk *Contextual Retrieval* on a larger multi-domain corpus remains valuable future work. The slight Hit@10 reduction reflects sibling-chunk competition in the lower top-10 ranks, which could be addressed via document-level deduplication or diversification rerankers. Generation results stem from a single greedy pass, so minor variations across prompting configurations should be read with care, and all generation metrics are reference-based rather than direct context-faithfulness audits.

Future work will extend ViCH-QA to other cultural domains, incorporate LLM-as-a-judge evaluation of factual precision, ablate summary lengths and generator models, and evaluate SACR within compiled inverted-index search engines such as Lucene or Tantivy.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '09*, 758–759. https://doi.org/10.1145/1571941.1572114

Do, P., Phan, T. H. V., & Gupta, B. B. (2021). Developing a Vietnamese Tourism Question Answering System Using Knowledge Graph and Deep Learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, *20*(5), 1-18. https://doi.org/10.1145/3453651

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, *55*(12), 1-38. https://doi.org/10.1145/3571730

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient Memory Management for Large Language Model Serving with PagedAttention. *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles* (pp. 611 - 6). https://doi.org/10.1145/3600006.3613165

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., & others. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 74–81.

Nguyen, H.-D., Dang, M.-A., Le, M.-T., & Le, M.-T. (2025). VietMEAgent: Culturally-Aware Few-Shot Multimodal Explanation for Vietnamese Visual Question Answering. *arXiv Preprint arXiv:2511.09058*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

Peoples, S. (2014). Intangible heritage and the museum: New perspective on cultural preservation. *International Journal of Heritage Studies*, *20*(5), 579–581. https://doi.org/10.1080/13527258.2014.913343

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Sun, J., Shi, J., Xie, Y., Wang, S., Parekh, J. R., Jiang, P., Shi, Z., Fan, J., Zheng, Q., Li, P., Wang, S., Liu, G., & Han, J. (2026). *Retrieval is Cheap, Show Me the Code: Executable Multi-Hop Reasoning for Retrieval-Augmented Generation*.

Tao, W., Xing, X., Li, Z., & Xu, X. (2025). SAKI-RAG: Mitigating Context Fragmentation in Long-Document RAG via Sentence-level Attention Knowledge Integration. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 1195–1213). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.63

Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2023). Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 10014–10037). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.557

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., & others. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://doi.org/10.52202/068431-1800

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. *International Conference on Learning Representations*.