

# An Empirical Study of Evidence Retrieval for Retrieval Augmented Generation over the Thai Royal Gazette

Pantai Suyasri<sup>1</sup> Pawitsapak Akarajardwong<sup>2</sup> Chompakorn Chaksangchaichot<sup>2</sup>

<sup>1</sup>School of Computing, KAIST, Daejeon, South Korea

<sup>2</sup>VISAI AI, Bangkok, Thailand

pantaisuyasri@kaist.ac.kr

{pawitsapaka\_visai, chompakornc\_pro}@vistec.ac.th

## Abstract

Retrieval-augmented generation depends on both the retrieval method and the units over which retrieval operates. Yet retrieval-unit choices remain underexplored for Thai legal documents, where language and document structure challenge common chunking assumptions. We introduce a 2,000-question Thai Royal Gazette dataset with single-hop and multi-hop tasks. The dataset annotates gold evidence as document character spans, which permits unit-independent evaluation. We compare 19 configurations across chunk sizes, boundary strategies, contextual chunks, atomic facts, whole documents, and retrieval methods. We evaluate them with span recall and precision, an Evidence-Fit judge, and pairwise judgments of answers generated from retrieved evidence. Contextual 200-word chunks with hybrid retrieval provide the strongest practical balance of evidence coverage, answer quality, and query cost. Across configurations, downstream scores correlate strongly with span recall, while span precision shows little association. These results favor retrieval designs that first recover complete evidence, then control cost and localization.

## 1 Introduction

Retrieval-Augmented Generation (RAG) grounds language model outputs in evidence retrieved from an external corpus (Lewis et al., 2020). Because the generator can only work with the evidence it receives, the evidence-retrieval stage sets the ceiling for the whole system, and building it requires two largely independent decisions. The first is what units the system ranks, such as fixed-size chunks, contextualized chunks, atomic propositions, or whole documents. Fixed-length embeddings lose semantic information as input length

grows (Luan et al., 2021), so smaller units reduce compression loss but can strip necessary context or split evidence across boundaries (Günther et al., 2025; Chen et al., 2024b). The second is how the system scores the units, by lexical term overlap, by similarity in a learned embedding space (Karpukhin et al., 2020), or by a fusion of the two, optionally re-scoring the top candidates with a more accurate but costlier second stage (Nogueira and Cho, 2020). A practical system must fix both decisions at once.

Existing findings on retrieval units and methods come mainly from English or general-domain corpora, and even there, different tasks and corpora favor different setups (Bhat et al., 2025). Thai adds its own complications. It is written without word delimiters, and English-centric tokenizers inflate token counts, which makes fixed-size segmentation less reliable (Petrov et al., 2023). This raises the practical question of whether techniques developed largely around English text, such as sentence-aligned boundaries, chunk overlap, contextualized chunks, and proposition extraction, transfer to a Thai legal corpus with different linguistic and structural properties.

Recent work has begun to fill this gap for legal documents. NitiBench demonstrated that segmentation choices substantially affect RAG performance on Thai statutes (Akarajardwong et al., 2025). Similarly, Li et al. (2023) decomposed long legal cases into *Facts*, *Court Reasoning*, and *Final Decision* sections to structure case retrieval. However, evidence retrieval over Thailand’s Royal Gazette has not been studied in depth. The Gazette is the state’s official publication of record for legislation, regulations, agency notifications, and court orders, spanning the breadth of Thai legal and administrative decisions. Each notice type has its own structure, which creates distinct evidence-retrieval challenges.

This paper systematically evaluates how retrieval-unit construction affects evidence retrieval over Thai Royal Gazette documents. We compare multi-

<sup>1</sup>This work was done as a part of Internship Program at VISAI.AI

ple strategies using a synthetic question–evidence dataset constructed from the target corpus. Our contributions include: (1) a comparison of chunking strategies for Thai Royal Gazette retrieval, (2) a dataset covering single- and multi-hop questions,<sup>1</sup> and (3) empirical recommendations for configuring evidence retrieval over Thai Royal Gazette documents.

## 2 Related Work

**Methods for Evidence Retrieval** Early evidence-retrieval systems used lexical matching, and BM25 remains a strong baseline for scoring weighted term overlap over an inverted index (Robertson and Zaragoza, 2009). However, lexical methods struggle when relevant passages express the same meaning using different terms. Dense retrieval addresses this problem by encoding the query and passage into dense vectors and comparing their similarity (Karpukhin et al., 2020). Dense and sparse methods can also be combined in a hybrid approach (Luan et al., 2021).

Two other families trade cost for accuracy. A reranker evaluates the top candidates from a cheaper first stage using a more accurate model that is too expensive to run on the entire corpus (Nogueira and Cho, 2020). Multivector methods such as ColBERT (Khattab and Zaharia, 2020) represent each query and passage as a set of per-token contextualized embeddings instead of a single pooled vector. Relevance is computed by comparing every query-token embedding against all passage-token embeddings, taking the maximum similarity for each query token, and summing these over the query.

**Units for Evidence Retrieval in RAG** Orthogonal to the retrieval method is the choice of retrieval unit. Because evidence retrieval matches a query to passages with similar semantic content, chunk size and construction can affect performance. Empirical work shows that no single chunk size is optimal across tasks, corpora, and embedding models (Bhat et al., 2025). Boundary placement also matters. Fixed-length windows can split sentences or separate related context, whereas sentence-aware chunking preserves linguistic units and can improve evidence-retrieval quality. Overlap between adjacent chunks can further reduce boundary loss by

placing content near a split in more than one retrieval unit (Stäbler et al., 2025).

Other approaches change the representation of the retrieval unit itself. Chen et al. (2024b) embed atomic propositions that are rewritten to be interpretable and unambiguous without surrounding context, making each unit more directly matchable to a query. Anthropic (2024) instead prepends contextual summaries to chunks so that otherwise missing document-level information remains available for the embedding. These methods can improve evidence retrieval in specific settings but add pre-processing costs. Existing work therefore shows that retrieval-unit design is consequential but task-dependent. However, these findings come mainly from English-language or general-domain datasets, so their applicability to Thai legal documents remains unclear.

**Synthetic Data for Evidence Retrieval** Evaluating evidence retrieval requires queries paired with their relevant passages, but manual annotation is costly. Large language models can generate synthetic questions from a corpus and ground them in known evidence (Liu et al., 2025).

This approach is scalable but introduces several issues. LLM-generated questions often exhibit limited linguistic and semantic diversity, converging on similar formulations even across models (Jiang et al., 2025). Individual items may also contain insufficient supporting evidence or purported multi-hop reasoning that does not actually require all designated passages. At the dataset level, synthetic queries may poorly approximate the intent and terminology distribution of real users (Zhang et al., 2026). Because of these limitations, generated items should be filtered, combining model-based checks for factors such as answerability from evidence with manual auditing of a sample (Zhang et al., 2026).

## 3 Methodology

### 3.1 Evidence-Retrieval Task

This work studies the *evidence retrieval* task: given a question  $q$  and a fixed corpus  $\mathcal{D}$ , a system ranks retrieval units drawn from  $\mathcal{D}$ . Each unit is associated with a source document and a span in that document’s text.

For each question, the dataset provides a set of gold character spans containing the evidence required to answer the question denoted by  $G_q =$

<sup>1</sup>The dataset is released at <https://huggingface.co/datasets/VISAI-AI/thai-gazette-evidence-retrieval> under CC BY 4.0. The hosting account is anonymized for review.

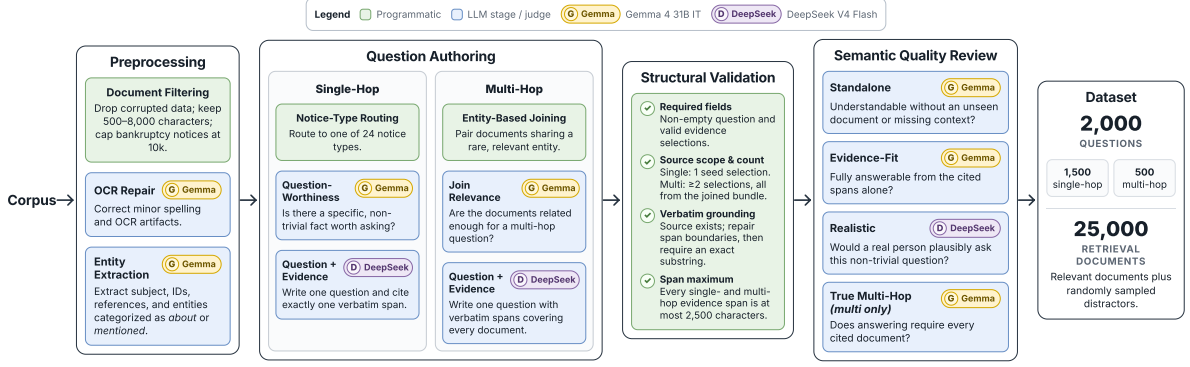


Figure 1: Dataset generation pipeline. Preprocessing cleans, OCR-repairs, and entity-tags each document, after which single-hop and multi-hop authoring produce candidate items. Every item must pass structural validation and then semantic quality review before it enters the dataset.

$\{(d_i, s_i, e_i)\}$ , where  $d_i$  identifies the source document and  $s_i$  and  $e_i$  are the start and end character offsets of the span within it. A system receives only  $q$  and the corpus and returns the top- $k$  ranked units, which are mapped to their source-document spans and evaluated by how completely and precisely they recover  $G_q$ .

To support this evaluation, we construct a Thai question-evidence dataset in which every question is paired with verbatim gold spans in one or more Royal Gazette documents. Because gold evidence is annotated in document character coordinates rather than as strategy-specific unit identifiers, we can compare different retrieval-unit constructions and retrieval methods against the same ground truth. Appendix B.4 shows one complete item of each question type.

### 3.2 Corpus

Thailand’s Royal Gazette is the state’s official publication of record for legislation, regulations, agency notifications, and court orders. We use the machine-readable release by [Open Law Data Thailand \(2025\)](#). We remove processing failures and length outliers, and cap formulaic notices, yielding 110,733 source documents<sup>2</sup>.

Because subsequent stages of the data-generation pipeline depend on exact character matching, we then apply a constrained LLM pass instructed to repair OCR errors and noise without changing content. Its output is the canonical text used for authoring and evaluation (Appendix A.2).

<sup>2</sup>More details on the dataset construction and examples are in Appendix A

### 3.3 Dataset Construction

Our comparison requires evidence annotations that stay invariant across retrieval-unit constructions, so we author questions from source documents and annotate their support as document-level character spans (§3.1). Authoring from known evidence produces these annotations at the scale and document diversity the comparison needs, which manual annotation could not reach under our budget. Because synthetic generation introduces the failure modes discussed in §2, generated data passes the validation pipeline in Figure 1.

**Single-hop authoring** We use *Notice-Type Routing* to assign each document to one of 24 Gazette notice types (Appendix B.1) and apply a prompt tailored to the information most salient to that type. Compared with a single generic prompt, which tended to produce similar questions across documents, type-specific prompts encourage coverage of the attributes that distinguish each notice type (Dai et al., 2022). A *Question-Worthiness* check then screens for a specific, non-trivial fact. Finally, an LLM writes one Thai *question* and cites one verbatim *evidence* span.

**Multi-hop authoring** An LLM entity-extraction pass over the corpus normalizes surface variants to a canonical form. Following Welbl et al. (2018), *Entity-Based Joining* then pairs documents that share one normalized bridge entity, restricted to rare entities that appear in two to eight documents. The bridge must be a principal subject of at least one paired document rather than an incidental mention, and the ranking prefers bridges central to both. Then, a *Join-Relevance* check screens whether the

pair is related enough to support a multi-hop question. Following the bridge-entity design of Yang et al. (2018), an LLM finally writes one *question* that combines a distinct fact from each document and cites verbatim *evidence* spans from every document used.

**Validation** We validate with defense-in-depth, as no single check catches every failure mode. *Structural Validation* check enforces the required number of source documents, maximum span length, and that the selected gold evidence passage exists verbatim in the text. Separate LLM judges then rate four semantic properties, applying large language models as reference-free evaluators in the manner of Liu et al. (2023). *STANDALONE* asks whether the question can be understood on its own, without unstated authoring context and without deictic references. *EVIDENCE-FIT* asks whether the cited span actually supports the stated answer. *REALISTIC* asks whether a person would plausibly ask this question of this document. *TRUE MULTI-HOP*, applied only to multi-hop items, asks whether every cited document is necessary, separating genuine bridge reasoning from questions one document already answers. During development, we iteratively spot-checked judge ratings against human assessments on small purposive samples and adjusted prompts or models when systematic disagreements appeared. As a final check on the pipeline, we also manually reviewed a sample, following the recommendation of Zhang et al. (2026). An author read 100 items. 95 of the 100 items were judged acceptable. The 5 remaining items contains evidence spans that partially fail to answer the whole question and were fixed manually.

## 4 Experimental Setup

### 4.1 Evaluation Dataset

All configurations use the same dataset from Section 3.3, with 2,000 questions (1,500 single-hop, 500 multi-hop) over a 25,000-document corpus holding every cited evidence document plus randomly sampled Gazette distractors (Appendix B).

### 4.2 Experimental Design

To evaluate evidence-retrieval performance across chunking strategies and retrieval models, we organize the analysis along two dimensions: the **retrieval-unit axis** and the **retrieval-method axis**.

**Retrieval-unit axis** We first compare fixed-size chunks of 50, 100, 200, and 400 words segmented with the dictionary-based newmm tokenizer. We retrieve  $k=16,8,4$ , and 2 chunks, respectively, so that each configuration has the same nominal retrieval budget of 800 words. For the remaining experiments, we use 200-word chunks. The recall gain from larger chunks tapers at this size (Table 1), and retrieving  $k = 4$  units leaves slack to cover both documents of a multi-hop item, whereas  $k = 2$  must place both of its units correctly.

We also evaluate two boundary heuristics as alternatives to standard 200-word chunks. The first adds a 50-word overlap between adjacent chunks. The second uses sentence-aligned chunks, in which each unit contains as many complete sentences as possible without exceeding 200 words and boundaries fall only at sentence endings. This prevents sentences from being split across retrieval units.

In addition, we test 200-word contextualized chunks following Anthropic (2024). We separately adapt the proposition-based retrieval approach of Chen et al. (2024b) by prompting an LLM to extract atomic propositions and retrieving  $k = 16$  propositions. While this does not exactly match the 800-word budget of previous experiments, we hypothesize that extracted propositions will be denser in semantics, making them roughly equivalent in content to the other conditions. Every budget quoted here is nominal, because units at document ends are short, so the text each condition actually delivers differs.

**Retrieval-method axis** The reference condition uses 200-word chunks, BGE-M3 embeddings (Chen et al., 2024a), dense retrieval, and  $k = 4$ . Holding the unit fixed, we compare dense retrieval with BM25 hybrid fusion, hybrid fusion followed by reranking with Qwen3-Reranker-4B (Zhang et al., 2025), and BGE-M3 multivector retrieval. We repeat the dense, hybrid, and reranked comparison over contextual and atomic-fact units. Whole-document units are compared with dense and multivector retrieval. Embedding ablations replace BGE-M3 with Voyage-4-Nano (Voyage AI, 2025) or Qwen3-Embedding-4B.

### 4.3 Retrieval Implementation

**Dense retrieval** Document and query vectors are L2 normalized and searched with an exact FAISS inner-product index (Douze et al., 2026).

**Hybrid retrieval and reranking** Thai is written without word delimiters, so BM25 tokenizes with newmm (Phatthiyaphaibun et al., 2023), the segmenter that also defines our word units (Appendix C). We then normalize the dense and sparse scores and add them for units retrieved by both methods. Score-based fusion preserves information about retrieval confidence (Fox and Shaw, 1993; Lee, 1997), unlike the reciprocal rank fusion of Cormack et al. (2009), which has been shown to perform worse (Bruch et al., 2023). Without reranking, each branch retrieves  $k$  units before fusion. With reranking, each branch retrieves  $2k$ , fusion retains a  $4k$  pool, and Qwen3-Reranker-4B returns the final  $k$ .

**Multivector retrieval** Exhaustive token-pair scoring is impractical at corpus scale, so BGE-M3 token vectors are indexed with PLAID (Santhanam et al., 2022), which quantizes them against a centroid codebook and prunes candidates on centroid interaction before any full scoring. We use two-bit compression, eight probed cells, and at most 8,192 full scores (Appendix D). Chunk-level and whole-document conditions both return  $k = 4$  units.

#### 4.4 Evaluation Metrics

For each question  $q$  and document  $d$ , let  $G_{q,d}$  denote the union of all gold character spans in  $d$  associated with  $q$ , and let  $R_{q,d}$  denote the union of retrieved character spans in  $d$  from  $q$ . Let  $\mathcal{D}_q^{\text{gold}}$  denote the set of gold documents for  $q$ , and let  $\mathcal{D}$  denote the full document collection. We merge overlapping spans within each document before computing their lengths. For any span set,  $|\cdot|$  denotes the total number of covered characters after merging overlaps. We then define the gold length  $g_q$ , overlap length  $o_q$ , and total retrieved length  $r_q^{\text{all}}$  as follows:

$$g_q = \sum_{d \in \mathcal{D}_q^{\text{gold}}} |G_{q,d}|, \quad (1)$$

$$o_q = \sum_{d \in \mathcal{D}_q^{\text{gold}}} |G_{q,d} \cap R_{q,d}|, \quad (2)$$

$$r_q^{\text{all}} = \sum_{d \in \mathcal{D}} |R_{q,d}|. \quad (3)$$

We define span recall and span precision as follows:

$$\text{SpanRec}(q) = \frac{o_q}{g_q}, \quad (4)$$

$$\text{SpanPrec}(q) = \frac{o_q}{r_q^{\text{all}}}. \quad (5)$$

SpanRec directly measures how well the chunking strategy recalls evidence. Span precision (SpanPrec) divides by everything retrieved, distractor documents included, so it measures how much of the delivered text is evidence and penalizes retrieving the wrong documents.

Span overlap does not establish that the retrieved text suffices to answer a question. We therefore also report the EVIDENCE-FIT judge from dataset review (Section 3.3, **Validation**), a Likert rating from 1 to 5, applied to a scenario-stratified sample of 100 questions. The same Gemma judge scores the same question identifiers for every configuration.

**Downstream generation** To further examine the effect of different configurations on downstream generation, we select 200 questions, and a Gemma 4 31B generator answers each question using only the retrieved text, employing the same concise prompt and deterministic decoding via zero temperature and fixed seed. Following the blinded pairwise protocol of Zheng et al. (2023), a Claude Opus 5 judge receives the question, gold evidence, and two anonymous answers, then selects either answer or a tie. We aggregate the pairwise judgments into an Elo rating for each configuration.

## 5 Results

The experimental results were presented in Table 1.

### 5.1 Retrieval units

**Larger chunks improve evidence coverage and judged retrieval quality.** Their higher precision partly reflects shorter delivered text, because document boundaries truncate more chunks. Even with less text, larger chunks preserve related context during indexing and answer generation. Whole-document retrieval extends this pattern and achieves the highest judge score among retrieval units. However, it is not a controlled chunk-size comparison. We use  $k=4$ , so the system can retrieve evidence from multiple documents for multi-hop questions. This choice increases the delivered-text budget, so its recall and precision are not directly comparable with the fixed-budget chunk conditions (Table 5).

| (a) Retrieval unit ( <i>single-vector dense</i> ) |              |              |             | (b) Retrieval method and encoder                    |              |              |             |
|---------------------------------------------------|--------------|--------------|-------------|-----------------------------------------------------|--------------|--------------|-------------|
| Unit                                              | SpanRec      | SpanPrec     | EvFit       | Method                                              | SpanRec      | SpanPrec     | EvFit       |
| <i>Word chunks, fixed 800-word budget</i>         |              |              |             | <i>Word 200w</i>                                    |              |              |             |
| 50w ( $k=16$ )                                    | 0.589        | 0.034        | 3.59        | vector*                                             | 0.765        | 0.057        | 3.94        |
| 100w ( $k=8$ )                                    | 0.677        | 0.045        | 3.73        | hybrid                                              | 0.818        | 0.059        | 4.30        |
| 200w ( $k=4$ )*                                   | 0.765        | 0.057        | 3.94        | hybrid + qwen rerank                                | 0.871        | 0.062        | 4.33        |
| 400w ( $k=2$ )                                    | 0.785        | <b>0.088</b> | 4.26        | multivector (ColBERT)                               | 0.800        | 0.057        | 4.02        |
| <i>Boundary ablations (200w)</i>                  |              |              |             | <i>Contextual 200w</i>                              |              |              |             |
| + overlap-50                                      | 0.786        | 0.060        | <u>4.28</u> | vector                                              | 0.823        | 0.062        | 4.28        |
| + sentence boundaries                             | 0.752        | 0.057        | 4.00        | hybrid                                              | 0.884        | <u>0.066</u> | <b>4.67</b> |
|                                                   |              |              |             | hybrid + qwen rerank                                | <b>0.925</b> | <b>0.067</b> | <u>4.63</u> |
| <i>Alternative representations</i>                |              |              |             | <i>Atomic facts (<math>k=16</math>)<sup>‡</sup></i> |              |              |             |
| contextual (200w)                                 | <u>0.823</u> | <u>0.062</u> | <u>4.28</u> | vector                                              | 0.647        | 0.068        | 4.25        |
| atomic facts ( $k=16$ ) <sup>‡</sup>              | 0.647        | 0.068        | 4.25        | hybrid                                              | 0.633        | 0.067        | 3.95        |
| whole document ( $k=4$ )                          | <b>0.863</b> | 0.043        | <b>4.44</b> | hybrid + qwen rerank                                | 0.694        | 0.072        | 4.17        |
|                                                   |              |              |             | <i>Whole document (<math>k=4</math>)</i>            |              |              |             |
|                                                   |              |              |             | dense                                               | 0.863        | 0.043        | 4.44        |
|                                                   |              |              |             | multivector (ColBERT)                               | <u>0.901</u> | 0.045        | 4.51        |
|                                                   |              |              |             | <i>Encoder ablations (word 200w, vector)</i>        |              |              |             |
|                                                   |              |              |             | voyage-4-nano                                       | 0.805        | 0.060        | 4.27        |
|                                                   |              |              |             | qwen3-4b                                            | 0.756        | 0.056        | 3.89        |

Table 1: Evidence-retrieval quality by unit and method, all with BGE-M3 unless noted. Higher is better. **EvFit** is the EVIDENCE-FIT rating from 1 to 5 (§3.3) on one stratified 100-question sample. \*Shared 200-word dense reference in both panels. Delivered text differs by row, so span metrics rank methods within a unit rather than across units (Table 5). <sup>‡</sup>Atomic-fact span metrics use the short cited spans, which inflates SpanPrec and caps SpanRec (Appendix C); these rows are therefore excluded from the span-metric highlighting. **Bold** marks the best and underline the second-best value in each column within a panel.

**Boundary modifications did not consistently improve on ordinary 200-word chunks.** Adding a 50-word overlap raised span recall slightly and the judge score more clearly. Aligning chunks to sentence boundaries instead lowered recall and left the judge score unchanged within noise. Overlap repairs some evidence split across adjacent chunks, which accounts for the small gain it produces, while sentence alignment with this corpus gives up control of the word budget without buying anything back.

**Contextualized chunks improve performance at additional preprocessing cost.** <sup>3</sup> Prepending a short document-level summary to each 200-word passage increased span recall, span precision, and the judge score. The summary gives the encoder document-level context for passages that otherwise resemble unattributed statutory text. Because only the original passage is scored by the judge and span metrics (Appendix C), the gains reflect better retrieval rather than added answer context.

<sup>3</sup>Contextual units require one full LLM pass over the corpus before indexing. This cost scales with corpus size but does not recur at query time.

**Atomic propositions trade evidence coverage for compactness.** Decomposing each document into standalone facts gives the highest span precision and the lowest span recall among the alternative representations, below the ordinary 200-word reference. Part of this precision advantage is definitional rather than substantive (§6, Appendix C). Propositions match contextual chunks on the judge score while delivering about three quarters as much text (Table 5). They also achieve the highest multi-hop judge score among single-vector units (Table 2). This decomposition isolates answer-bearing statements but removes surrounding text that helps readers verify them.

## 5.2 Retrieval methods

**Hybrid retrieval improves performance for word and contextual units but lowers it for atomic propositions.** The likely reason is that lexical matching operates on surface forms, and propositions are LLM-written restatements in which the original strings a BM25 index would match no longer appear. Qwen reranking added a further 0.04 to 0.06 SpanRec to each of the three representations, giving contextual chunks the highest SpanRec we measured. Because the reranking gain holds across

representations, it transfers from one unit to another, and the two axes can be tuned separately. Reranking costs two to three orders of magnitude more query time than the fusion step it reranks (Table 7).

**Encoder and multivector changes produced small gains.** On 200-word chunks, Voyage-4-Nano and BGE-M3 multivector retrieval increased SpanRec, while Qwen3-Embedding-4B did not improve over the baseline (Appendix E.2). Multivector retrieval produced similar SpanRec gains for chunks and whole documents. Thus, late interaction did not especially benefit longer units. EvFit improved significantly only for Voyage-4-Nano. However, multivector retrieval increased query time by about two orders of magnitude (Table 7).

### 5.3 Multi-hop behavior

Table 2 splits every condition into its single-hop and multi-hop questions. Multi-hop retrieval is a harder task because it must recover evidence for both hops. Therefore, EvFit is lower for every configuration, and multi-hop span recall trails single-hop in all 19 conditions. Multi-hop span precision exceeds single-hop in every row, because a multi-hop question carries more gold evidence, so a fixed 800-word budget lands a larger share of its text inside gold documents.

**Method changes narrow the multi-hop deficit.** On ordinary 200-word chunks, multi-hop span recall trails single-hop by 0.050 under dense retrieval. The deficit narrows under hybrid, reranked, and multivector retrieval, and it nearly closes under Voyage-4-Nano. A single dense query vector averages toward one of the two documents a multi-hop item needs, whereas lexical and late-interaction signals can recover the second on a rare name or identifier.

### 5.4 Downstream answer quality

**Contextual 200-word hybrid retrieval leads the downstream ranking.** It is ordered above its Qwen-reranked variant, reranked atomic propositions, and document multivector retrieval, but all four fall within 0.03 match score of each other. Ties account for 62.2% of all pairwise comparisons, and the two contextual hybrid variants draw eight of their ten head-to-head games. Whole-document systems also rank highly.

**Downstream answer quality tracks evidence coverage more closely than retrieval precision.**

Across 19 configurations, downstream scores correlated with SpanRec at Spearman  $\rho = 0.688$  and with SpanPrec at 0.293. Contextual chunks, whole documents, and hybrid methods ranked highest downstream. Atomic facts achieved the highest span precision and placed 3rd, 11th, and 13th, so precision-heavy units are not penalized downstream, but coverage separates the ranking more reliably.

**The downstream spread is narrow.** Retrieval-unit choice moves retrieval quality far more than it moves answer quality. Span recall spans nearly 0.3 across our configurations, while downstream match scores stay clustered near even, from 0.340 to 0.579. Ties account for 62.2% of pairwise comparisons, so the judge rarely separates two configurations at all.

## 6 Discussion

**Measuring span-level grounding across retrieval units is an open problem.** When the unit type and final retrieval depth are fixed, each method is scored against the same gold spans over the same candidate units. Differences therefore reflect the retrieval method. Across unit types, however, scores also depend on how each unit partitions and delivers the source text.

Unit-level rank metrics require each gold span to be mapped to every overlapping retrieval unit. This mapping changes with the unit construction. A coarse unit may contain a complete gold span, while a fine unit divides the same span across several units. Binary relevance treats partial and complete overlap alike, so rank scores can change even when the retrieved evidence is similar. We therefore use span metrics instead of unit-level rank metrics.

Span metrics score character overlap against fixed gold offsets and avoid this mapping. However, the retrieval unit still affects these metrics. Coarse units are truncated at document boundaries more often, which can reduce delivered text and raise SpanPrec. Atomic propositions also do not tile the source text, which limits their attainable SpanRec.

We compare retrieval units through blinded pairwise comparisons of generated answers. The judge sees answers rather than retrieval units, so retrieval-unit boundaries do not affect this comparison directly. As an alternative that does not require answer generation, pointwise judgments of retrieved evidence also tracked the downstream ranking at lower

(a) Retrieval unit (single-vector dense)

| Unit                                      | SpanRec      |              | SpanPrec     |              | EvFit       |             |
|-------------------------------------------|--------------|--------------|--------------|--------------|-------------|-------------|
|                                           | SH           | MH           | SH           | MH           | SH          | MH          |
| <i>Word chunks, fixed 800-word budget</i> |              |              |              |              |             |             |
| 50w ( $k=16$ )                            | 0.593        | 0.578        | 0.026        | 0.060        | 3.81        | 2.92        |
| 100w ( $k=8$ )                            | 0.680        | 0.669        | 0.033        | 0.082        | 4.08        | 2.68        |
| 200w ( $k=4$ )*                           | 0.777        | 0.727        | 0.042        | 0.099        | 4.25        | 3.00        |
| 400w ( $k=2$ )                            | 0.817        | 0.689        | <b>0.068</b> | <b>0.148</b> | 4.56        | 3.36        |
| <i>Boundary ablations (200w)</i>          |              |              |              |              |             |             |
| + overlap-50                              | 0.800        | 0.745        | 0.045        | 0.105        | <u>4.63</u> | 3.24        |
| + sentence boundaries                     | 0.762        | 0.722        | 0.043        | 0.101        | 4.29        | 3.12        |
| <i>Alternative representations</i>        |              |              |              |              |             |             |
| contextual (200w)                         | <u>0.842</u> | <u>0.768</u> | <u>0.046</u> | <u>0.109</u> | 4.56        | 3.44        |
| atomic facts ( $k=16$ ) <sup>‡</sup>      | 0.675        | 0.564        | 0.052        | 0.114        | 4.41        | <b>3.76</b> |
| whole document ( $k=4$ )                  | <b>0.880</b> | <b>0.814</b> | 0.031        | 0.078        | <b>4.68</b> | <u>3.72</u> |

(b) Retrieval method and encoder

| Method                                              | SpanRec      |              | SpanPrec     |              | EvFit       |             |
|-----------------------------------------------------|--------------|--------------|--------------|--------------|-------------|-------------|
|                                                     | SH           | MH           | SH           | MH           | SH          | MH          |
| <i>Word 200w</i>                                    |              |              |              |              |             |             |
| vector*                                             | 0.777        | 0.727        | 0.042        | 0.099        | 4.25        | 3.00        |
| hybrid                                              | 0.824        | 0.801        | 0.043        | 0.105        | 4.44        | 3.88        |
| hybrid + qwen rerank                                | 0.875        | <u>0.857</u> | 0.045        | 0.113        | 4.48        | 3.88        |
| multivector (ColBERT)                               | 0.809        | 0.774        | 0.042        | 0.102        | 4.21        | 3.44        |
| <i>Contextual 200w</i>                              |              |              |              |              |             |             |
| vector                                              | 0.842        | 0.768        | 0.046        | 0.109        | 4.56        | 3.44        |
| hybrid                                              | 0.893        | 0.855        | <u>0.048</u> | <u>0.118</u> | <b>4.89</b> | <u>4.00</u> |
| hybrid + qwen rerank                                | <b>0.936</b> | <b>0.891</b> | <b>0.049</b> | <b>0.119</b> | <u>4.83</u> | <b>4.04</b> |
| <i>Atomic facts (<math>k=16</math>)<sup>‡</sup></i> |              |              |              |              |             |             |
| vector                                              | 0.675        | 0.564        | 0.052        | 0.114        | 4.41        | 3.76        |
| hybrid                                              | 0.654        | 0.571        | 0.051        | 0.115        | 4.09        | 3.52        |
| hybrid + qwen rerank                                | 0.720        | 0.614        | 0.055        | 0.125        | 4.31        | 3.76        |
| <i>Whole document (<math>k=4</math>)</i>            |              |              |              |              |             |             |
| dense                                               | 0.880        | 0.814        | 0.031        | 0.078        | 4.68        | 3.72        |
| multivector (ColBERT)                               | <u>0.919</u> | 0.849        | 0.033        | 0.082        | 4.73        | 3.84        |
| <i>Encoder ablations (word 200w, vector)</i>        |              |              |              |              |             |             |
| voyage-4-nano                                       | 0.805        | 0.804        | 0.044        | 0.109        | 4.43        | 3.80        |
| qwen3-4b                                            | 0.767        | 0.724        | 0.042        | 0.097        | 4.12        | 3.20        |

Table 2: Single-hop (SH) and multi-hop (MH) breakdown of Table 1, with the same panels, groups, and row labels. Span metrics use all 2,000 questions, 1,500 single-hop and 500 multi-hop. EvFit rests on 75 single-hop and 25 multi-hop ratings, so the MH column is indicative rather than precise. Bold and underline mark the best and second-best value per column within a panel, with the <sup>‡</sup> span-metric exclusion of Table 1.

| Rank | Configuration                 | Elo    | Score |
|------|-------------------------------|--------|-------|
| 1    | contextual 200w hybrid        | 1553.6 | .579  |
| 2    | contextual 200w hybrid + Qwen | 1545.0 | .564  |
| 3    | atomic hybrid + Qwen          | 1539.8 | .560  |
| 4    | document multivector          | 1537.0 | .555  |
| 5    | document dense                | 1529.3 | .542  |
| 6    | word 200w hybrid              | 1518.3 | .529  |
| 7    | word 200w Voyage embed        | 1514.2 | .518  |
| 8    | word 200w multivector         | 1510.4 | .513  |
| 9    | word 200w overlap dense       | 1509.8 | .513  |
| 10   | word 200w hybrid + Qwen       | 1503.3 | .505  |
| 11   | atomic dense                  | 1502.7 | .505  |
| 12   | contextual 200w dense         | 1500.4 | .500  |
| 13   | atomic hybrid                 | 1497.2 | .497  |
| 14   | sentence 200w dense           | 1496.3 | .495  |
| 15   | word 400w dense               | 1489.4 | .484  |
| 16   | word 200w dense               | 1472.1 | .458  |
| 17   | word 200w Qwen3 embed         | 1451.9 | .429  |
| 18   | word 100w dense               | 1442.0 | .413  |
| 19   | word 50w dense                | 1387.5 | .340  |

Table 3: Downstream ranking of all 19 configurations over 200 stratified questions, 189 to 190 games each. Claude Opus 5 judges every non-identical answer pair, and byte-identical answers are scored as ties without a model call. Score assigns one point to a win and one half to a tie.

cost, with a rank correlation of 0.774. However, both evaluations remain exposed to LLM judge bias.

### Coverage should take priority over precision.

The correlation pattern suggests that a large-context reader tolerates irrelevant retrieved text but cannot recover evidence that retrieval omits. Retrieval design should therefore prioritize coverage until the text budget binds, rather than trim distractors. The

narrow downstream spread partly reflects a dataset whose questions are often answerable from partial evidence, and partly a generator robust to imperfect context. The practical reading is that the difference between a reasonable configuration and the best tested one is modest, so the first-order decision is to avoid the clearly poor designs rather than to tune the last few points of recall.

**Practical recommendation.** Among the tested configurations, contextual 200-word chunks with hybrid retrieval give the strongest practical trade-off between evidence coverage, downstream performance, and query latency. When contextual preprocessing is unavailable, hybrid retrieval over ordinary 200-word chunks is the strongest tested low-cost alternative.

## 7 Conclusion

Retrieval-unit design and retrieval method act on Thai legal RAG through one shared channel, which is how much of the gold evidence reaches the generator. Units that preserve document context and methods that recover rare lexical forms both widen that channel, and downstream answer quality follows it rather than the precision of the retrieved text. Precision is also the quantity most sensitive to how a unit is shaped, so it separates systems reliably only when the unit is held fixed. Any study comparing chunking strategies across units inherits that

difficulty, and a unit-agnostic measure of whether retrieved text grounds an answer would remove it.

## Acknowledgement

This study was conducted as part of an internship project at VISAI. We thank Vidyasirimedhi Institute of Science and Technology (VISTEC) for the computational resources that made this study possible as well as their continued support and guidance throughout the study.

## Limitations

**Corpus** All results come from one Thai legal collection, so configuration rankings may not transfer to other jurisdictions, languages, or document genres. Its canonical text is LLM-repaired OCR output, and because Thai has no word delimiters, chunk boundaries, retrieval budgets, and BM25 terms all depend on a dictionary-based segmenter. The 800-word budget is nominal both across languages and within our own conditions, where 400-word chunks deliver 484 words per question against 767 for 50-word chunks (Table 5).

**Synthetic queries** The dataset supports controlled comparison over questions grounded in Gazette documents, but it does not reproduce the distribution of real user queries. Questions and gold spans are LLM-authored and LLM-filtered, not written by legal professionals. An author read 100 items and judged 95 acceptable, and in the five remaining items the cited span only partly answers the question. That single non-expert audit roughly estimates item-level error.

**Cross-unit comparability** Gold evidence stays fixed across retrieval units, so span metrics rank methods comparably within a unit. Across units the delivered text differs, atomic-fact and whole-document conditions are not budget matched, and atomic units depend on LLM extraction and span alignment. Their cited spans do not tile the document and cover 92.5% of the gold characters, capping atomic SpanRec at 0.925. We therefore read span metrics within comparable unit families and use downstream answer evaluation across them (Section 6).

**Model dependence** EvFit rests on one stratified 100-question sample, 25 items of it multi-hop, scored by a single Gemma judge, and its intervals resolve ten of our eighteen contrasts (Appendix E.2),

so we treat it as descriptive throughout. Downstream evaluation uses 200 questions, one generator, and one pairwise judge from a different model family, which avoids self-preference toward the generator’s own outputs. We did not validate that judge against human pairwise preferences. Ties are 62.2% of the 1,800 comparisons, and 563 of them are byte-identical answers recorded without a judge call. Contextual hybrid retrieval and its reranked variant record 1 win, 1 loss, and 8 ties against each other (Appendix E.3).

## References

- Pawitsapak Akarajardwong, Pirat Pothavorn, Chompakorn Chaksangchaichot, Panuthep Tasawong, Thitiwat Nopparatbundit, Keerakiat Pratai, and Sarana Nutanong. 2025. [NitiBench: Benchmarking LLM frameworks on Thai legal question answering capabilities](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34304–34327, Suzhou, China. Association for Computational Linguistics.
- Anthropic. 2024. Introducing contextual retrieval. <https://www.anthropic.com/engineering/contextual-retrieval>. Anthropic Engineering Blog.
- Sinchana Ramakanth Bhat, Max Rudat, Jannis Spiekermann, and Nicolas Flores-Herr. 2025. [Rethinking chunk size for long-document retrieval: A multi-dataset analysis](#). *Preprint*, arXiv:2505.21700.
- Sebastian Bruch, Siyu Gai, and Amir Ingber. 2023. [An analysis of fusion functions for hybrid retrieval](#). *ACM Trans. Inf. Syst.*, 42(1).
- Antoine Chaffin and Raphaël Sourty. 2025. [Pylate: Flexible training and retrieval for late interaction models](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM ’25*, page 6334–6339, New York, NY, USA. Association for Computing Machinery.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. [M3-Embedding: Multilinguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335. Association for Computational Linguistics.
- Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024b. [Dense X retrieval: What retrieval granularity should we use?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA. Association for Computational Linguistics.

- Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. [Reciprocal rank fusion outperforms condorcet and individual rank learning methods](#). In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA. Association for Computing Machinery.
- Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2022. [Promptagator: Few-shot dense retrieval from 8 examples](#). *Preprint*, arXiv:2209.11755.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2026. [The faiss library](#). *IEEE Transactions on Big Data*, 12(2):346–361.
- Edward A. Fox and Joseph A. Shaw. 1993. [Combination of multiple searches](#). In *Text Retrieval Conference*.
- Michael Günther, Isabelle Mohr, Daniel James Williams, Bo Wang, and Han Xiao. 2025. [Late chunking: Contextual chunk embeddings using long-context embedding models](#). *Preprint*, arXiv:2409.04701.
- Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. 2025. [Artificial hivemind: The open-ended homogeneity of language models \(and beyond\)](#). *Preprint*, arXiv:2510.22954.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Omar Khattab and Matei Zaharia. 2020. [ColBERT: Efficient and effective passage search via contextualized late interaction over BERT](#). In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48. ACM.
- Joon Ho Lee. 1997. [Analyses of multiple evidence combination](#). In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '97, page 267–276, New York, NY, USA. Association for Computing Machinery.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NeurIPS 2020, Red Hook, NY, USA. Curran Associates Inc.
- Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2023. [Lecardv2: A large-scale chinese legal case retrieval dataset](#). *Preprint*, arXiv:2310.17609.
- Wanlong Liu, Junying Chen, Ke Ji, Li Zhou, Wenyu Chen, and Benyou Wang. 2025. [RAG-instruct: Boosting LLMs with diverse retrieval-augmented instructions](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3865–3888, Suzhou, China. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. [Sparse, dense, and attentional representations for text retrieval](#). *Transactions of the Association for Computational Linguistics*, 9:329–345.
- Rodrigo Nogueira and Kyunghyun Cho. 2020. [Passage re-ranking with bert](#). *Preprint*, arXiv:1901.04085.
- Open Law Data Thailand. 2025. [Royal gazette thailand \(Ratchakitcha\) dataset](#). Hugging Face. Accessed 10 July 2026.
- Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. [Language model tokenizers introduce unfairness between languages](#). In *Advances in Neural Information Processing Systems*.
- Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiawat, Thanathip Suntornrtp, and Can Udomcharoenchaikit. 2023. [PyThaiNLP: Thai natural language processing in Python](#). In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 25–36. Association for Computational Linguistics.
- Stephen Robertson and Hugo Zaragoza. 2009. [The probabilistic relevance framework: BM25 and beyond](#). *Foundations and Trends in Information Retrieval*, 3(4):333–389.
- Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022. [Plaid: An efficient engine for late interaction retrieval](#). In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1747–1756, New York, NY, USA. Association for Computing Machinery.
- Maximilian Stäbler, Steffen Turnbull, Tobias Müller, Chris Langdon, Jorge Marx-Gómez, and Frank Köster. 2025. [The impact of chunking strategies on domain-specific information retrieval in rag systems](#).

In 2025 *IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–6.

Voyage AI. 2025. [voyage-4-nano](#). Hugging Face. Accessed 17 July 2026.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. [Constructing datasets for multi-hop reading comprehension across documents](#). *Transactions of the Association for Computational Linguistics*, 6:287–302.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Kaituo Zhang, Mingzhi Hu, Hoang Anh Duy Le, Faraha Kabir Torsha, Zhimeng Jiang, Minh Khai Bui, Chia-Yuan Chang, Yu-Neng Chuang, Zhen Xiong, Ying Lin, Guanchu Wang, and Na Zou. 2026. [A survey on evaluating quality and trustworthiness in LLM-generated data](#). *Preprint*, arXiv:2601.17717.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *Preprint*, arXiv:2506.05176.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.

## A Corpus Construction

### A.1 Corpus Selection and Filtering

The corpus builder pins revision 25572e1e18f20e1aa3e42e3ba1059ec5c0176f6a of the source dataset. It merges OCR and metadata rows by PDF filename and applies the following operations in order.

1. **Text recovery.** Prefer the formatted OCR output. If it is absent, concatenate valid page-level raw or Markdown outputs.
2. **Artifact removal.** Discard text containing one of the upstream processing-error markers encoded in the corpus builder.
3. **Script filter.** Discard documents whose ratio of ASCII characters to all characters is at least 0.30.

4. **Length filter.** Retain documents containing 500 through 8,000 characters, inclusive.

5. **Class cap.** Identify bankruptcy, incapacity, and receivership material with a regular expression and retain at most 10,000 matching documents.

6. **Shuffle.** Shuffle the retained rows with pseudorandom seed 42.

These operations yield 110,733 documents. All steps are deterministic and use no model.

### A.2 OCR Repair and Prompt

OCR repair runs Gemma 4 31B. Each request asks the model to repair only recognition errors within words, including displaced marks, confusable consonants, broken words, and stray characters. The prompt prohibits deletion, paraphrase, numeral changes, reordering, and layout changes, and instructs the model to preserve uncertain tokens. Requests use temperature 0 on the first attempt, permit two retries, and increase temperature to 0.3 only after a failed attempt.

The system prompt states the repair categories and the fidelity constraints.

You are an expert proofreader of OCR output from Thai Royal Gazette (ราชกิจจานุเบกษา) legal announcements. Your job is to repair OCR recognition errors so the text reads as the original document.

Work word by word. Fix only the characters that OCR mis-recognized inside each word:

- Misplaced or missing vowel marks (สระ) and tone marks (วรรณยุกต์).
- Visually similar consonants swapped (e.g. ค/ด, น/ม, พ/ฟ, ข/ช).
- Broken or merged words and stray ASCII characters inside Thai words.
- Wrong month names and obviously corrupted official terms (ราชกิจจานุเบกษา, พระราชบัญญัติ, มูลนิธิ, ฯ).

Hard constraints:

- Preserve EVERY word and clause. The corrected text must contain the same words in the same order as the input – never delete, shorten, merge, reorder, or summarize a phrase, and never replace a word with a synonym. You repair spelling, not wording.
- A garbled word still maps to one corrected word; do not collapse several words into fewer.
- Keep every numeral in its original script and value. Do not renumber anything.
- Keep the layout: line breaks, list numbering, spacing, and any Markdown markers (#, ##) exactly as they appear.
- If you cannot confidently repair a token by fixing characters, keep it exactly as-is rather than guessing or dropping it.
- Output ONLY the corrected text.

The accompanying user prompt.

Correct the OCR errors in this announcement and return the full corrected text:

```
{text}
```

## B Dataset Construction

### B.1 Single-Hop Construction

The taxonomy was not fixed in advance. We prompted Claude Code to inspect the entire source dataset and identify recurring Royal Gazette notice forms. Claude Code proposed an initial taxonomy, which we reviewed and consolidated into the following 24 notice types, in routing order:

1. Bankruptcy discharge
2. Bankruptcy judgment
3. Bankruptcy receivership
4. Incapacity order
5. Missing-person order
6. Association registration
7. Foundation registration
8. Association or foundation change
9. Registration revocation
10. Buddhist temple establishment
11. Catholic church recognition
12. Condominium registration
13. Partnership and company registration list
14. Industrial product standard
15. Appointment
16. Royal decoration or medal
17. Local ordinance
18. Land-title or replacement-title notice
19. Mining concession
20. Bank branch opening or relocation
21. Electoral district boundary
22. Labor union or employers' association
23. Correction
24. Rule or notification

*Notice Type Routing* uses no model. We wrote one regular-expression matcher for each notice type. The router searches up to three heading-like lines and the first 18 non-empty lines among the first 80 lines of each document. A document can match more than one type. In that case, the router selects the authoring type deterministically from the matches using seed 42 and a hash of the document identifier. The default typed-only queue excludes documents assigned to the unknown fallback class.

*Question-Worthiness Screening*, which runs Gemma 4 31B, and the author both see at most the first 12,000 characters. Screening retains scores of 4 or 5. *Question & Evidence Generation* runs DeepSeek V4 Flash with a class-specific prompt followed by shared constraints, temperature 0.4, a maximum of 1,800 output tokens, and at most two parsing retries. *Structural Validation* uses no model. It requires one non-empty question, exactly

one selected span, the seed document as its source, verbatim grounding after boundary repair, and a maximum span length of 2,500 characters.

### B.2 Multi-Hop Construction

*Entity Extraction* runs Gemma 4 31B at temperature 0 over the raw document text and emits five fields named *subject*, *about*, *mentioned*, *ids*, and *refs*. The discovery index treats *about* and *ids* as primary keys and the remaining entity lists as secondary keys. It normalizes Thai digits, whitespace, surrounding punctuation, and a short list of honorifics. Keys shorter than four characters are discarded.

*Entity-Based Document Joining* uses no model. An eligible bridge has document frequency from 2 through 8. Pairs for which the bridge is secondary on both sides are removed. Remaining pairs are ranked using the inverse document frequency of their best bridge, multiplied by 0.4 when the bridge is secondary on one side. The builder rejects same-issue, same-series, same-page duplicates and pairs sharing more than five entities. It permits each document in at most three bundles and caps each unordered document-category pair at 400 bundles. *Join Relevance Screening* runs Gemma 4 31B and requires a score of 4 or 5 before authoring.

During *Question & Evidence Generation*, the author is DeepSeek V4 Flash. It receives at most 10,000 characters per document and selects the minimal set of at least two source documents. *Structural Validation* uses no model. It restricts selections to the joined bundle, verifies each span verbatim after boundary repair, and applies the same 2,500-character maximum used for single-hop evidence. The TRUE MULTI-HOP check additionally reads the full selected documents when scoring whether all cited documents are necessary.

### B.3 Quality Review and Export

The judges are the semantic layer of the defense-in-depth described in §3.3. Structural validation can only confirm that a span is present and that the counts are right, so each remaining failure mode gets one judge.

**STANDALONE** Can the question be understood without seeing an unstated document, and does it name the entities and context needed to identify its target? This rejects deictic questions such as “this notice” whose meaning depends on hidden authoring context.

**EVIDENCE-FIT** Does the cited span support the stated answer? A verbatim span can be quoted correctly and still fail to license the answer, which would credit a retriever for finding text that does not settle the question.

**REALISTIC** Would a person plausibly ask this? The check reads the source excerpts as well, because a question can sound natural while resting on a role or a year that OCR noise invented.

**TRUE MULTI-HOP** Is every cited document necessary? This separates genuine bridge reasoning from items that one document already answers and that would otherwise inflate multi-hop difficulty.

Two screening stages also run before authoring and reject documents or bundles that offer nothing question-worthy, which keeps the authoring model from manufacturing questions out of thin material. The judge and screening prompts appear verbatim in Appendix B.5.3.

Default automated acceptance requires every applicable LLM-as-Judge check to return 4 or 5. The current calibrated routing assigns STANDALONE, EVIDENCE-FIT, and TRUE MULTI-HOP, together with both screening stages, to Gemma 4 31B. The REALISTIC check uses DeepSeek V4 Flash. The released provenance reflects the versioned construction process. All 1,500 single-hop items and 172 multi-hop items were judged with DeepSeek V4 Flash, while the 328 later multi-hop additions used the criterion-specific routing. Those additions also require a content-hash-bound manual approval. A manual override of an automated rejection is accepted only when the audit explicitly records the override and a non-empty justification.

At export, the implementation reruns the character-exact span check, derives the scenario label from the number of selected documents, and assigns a stable question identifier from the immutable generation seed. The *Evidence-Retrieval Corpus* includes every relevant document and adds distractors sampled with seed 42 until it reaches 25,000 documents. A distractor is ineligible when repetition removal would shrink its OCR text by more than 50%. The emitted *Dataset* contains 2,000 questions, with 1,500 single-hop and 500 multi-hop items.

```
qid          q7ddcdb65410c
question_type single_source

Question:
นายสุกิจ ศรีปิตต์ได้รับแต่งตั้งให้ดำรงตำแหน่งนายกสภามหาวิทยาลัยราชภัฏมหาสารคามตั้งแต่วันที่
[From what date was Mr. Suwukit Sripit appointed President of the University Council of Rajabhat Maha Sarakham University?]

Gold evidence: doc 2025-09-12-00081794.pdf,
chars [1233, 1373)
จึงมีพระบรมราชโองการโปรดเกล้าโปรดกระหม่อมแต่งตั้งบุคคล
ดังกล่าว ให้ดำรงตำแหน่ง
นายกสภามหาวิทยาลัยราชภัฏมหาสารคาม ตั้งแต่วันที่ ๖ กันยายน
๒๕๖๘
[There is accordingly a Royal Command appointing the said person as President of the University Council of Rajabhat Maha Sarakham University, effective from 6 September B.E. 2568 (2025).]
```

Figure 2: A released single-source item.

## B.4 Dataset Examples

Figures 2 and 3 show one released item of each question type, reproduced verbatim with English translations added in brackets. Character offsets are half-open intervals into the OCR text of the cited source document. The multi-source item follows the two-stage bankruptcy pattern, in which the receivership order and the bankruptcy judgment appear in separate notices, so neither document alone answers the question. Person and company names appear as the OCR produced them.

## B.5 Prompt Templates

### B.5.1 Single-Hop Authoring

Notice type routing is a deterministic regular-expression cascade and uses no model, so it has no prompt. *Question-Worthiness Screening* scores a candidate seed document.

```
You rate ONE Royal Gazette document as a
candidate for a single-hop retrieval dataset.

Question: does this document contain at least
one clean, specific, realistic Gazette fact,
rule,
status, event, duty, fee, penalty, appointment,
registration, cancellation, effective date, or
legal effect that could support ONE useful
single-hop question with ONE gold evidence
span?

Be selective. A good document has a non-trivial
thing a real user might search the Gazette to
verify, and enough clean text to quote as
evidence.

Reject or score low when the document is
mostly:
- boilerplate or administrative filler;
- a bulk roster/list where only arbitrary rows
  can be asked about;
- signer/registrar/case-number/Gazette-volume
  trivia;
- OCR-garbled text;
- title-only value where the answer is obvious
  without reading the document;
```

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>qid q507aa401bafb<br/>question_type multi_source</p> <p>Question:<br/>         ห้างหุ้นส่วนจำกัด ฟิสิษฐ์ เบสเตอร์ คดีหมายเลขแดงที่<br/>         ล.๓๒๐๓/๒๕๖๗<br/>         ศาลล้มละลายกลางมีคำสั่งพิทักษ์ทรัพย์เด็ดขาดเมื่อวันใด<br/>         และพิพากษาให้เป็นบุคคลล้มละลายเมื่อวันใด<br/>         [In red case number L.3213/2567, on what date<br/>         did the Central Bankruptcy<br/>         Court order absolute receivership of Phisit<br/>         Bester Limited Partnership,<br/>         and on what date did it adjudge it bankrupt?]</p> <p>Gold evidence 1: doc 2024-011569.pdf, chars<br/>         [199, 293)<br/>         ศาลได้มีคำสั่งลงวันที่ ๑๗ มิถุนายน ๒๕๖๗ ให้พิทักษ์ทรัพย์ของ<br/>         ห้างหุ้นส่วนจำกัด ฟิสิษฐ์ เบสเตอร์<br/>         [The court issued an order dated 17 June B.E.<br/>         2567 (2024) placing the<br/>         assets of Phisit Bester Limited Partnership<br/>         under receivership.]</p> <p>Gold evidence 2: doc 2025-01-22-00056384.pdf,<br/>         chars [156, 316)<br/>         ศาลล้มละลายกลางได้พิพากษาให้ ห้างหุ้นส่วนจำกัด ฟิสิษฐ์<br/>         เบสเตอร์ ลูกหนี้ที่ ๑ นายฟิสิษฐ์ จันทร์ ลูกหนี้ที่ ๒<br/>         เป็นบุคคลล้มละลายแล้ว ตั้งแต่วันที่ ๒๖ พฤศจิกายน ๒๕๖๗<br/>         [The Central Bankruptcy Court has adjudged<br/>         Phisit Bester Limited<br/>         Partnership, debtor no. 1, and Mr. Phisit Chan,<br/>         debtor no. 2, bankrupt,<br/>         effective from 26 November B.E. 2567 (2024).]</p> | <p>volume/page/date trivia, list-position<br/>         trivia, "what is this notice about", and<br/>         questions whose<br/>         answer is already leaked in the question.<br/>         - SOURCE NAMING: do not use<br/>         "ตามประกาศในราชกิจจานุเบกษา" as the source<br/>         identifier. Name the<br/>         actual law, notice, body, organization,<br/>         person, product, place, or program instead.<br/>         - VOICE: natural typed Thai, not legalese.<br/>         Arabic numerals are fine.<br/>         - Skip weak material: if the best available<br/>         evidence is too thin, too long, too generic, or<br/>         too<br/>         OCR-garbled, set `skip` to true and leave<br/>         `question` and `selected` empty.<br/>         - Do not write an answer. The dataset uses the<br/>         question and gold evidence span, not answer<br/>         authoring.</p> <p>Reply with ONE VALID JSON OBJECT ONLY. No<br/>         prose, no markdown, no code fence.<br/>         Always include exactly these keys: `skip`,<br/>         `reason`, `question`, `selected`.<br/>         `selected` must always be an array.</p> <p>Good output when keeping:<br/>         {"skip":false,"reason":"","question":"คำถามภา<br/>         ษาไทย","selected":[{"doc_id":"2012-000001.pdf"<br/>         }, {"evidence_span":"ข้อความหลักฐานแบบคัดลอกตรง"}]}</p> <p>Good output when skipping:<br/>         {"skip":true,"reason":"เอกสารไม่รองรับแบบคำดา<br/>         มที่เลือกพร้อมหลักฐานสะอาด","question":"","selecte<br/>         d":[]}</p> |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Figure 3: A released multi-source item.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>- private minutiae unlikely to be useful for<br/>         legal, administrative, journalistic, civic, or<br/>         compliance use.</p> <p>Rate from 1 to 5:<br/>         5 = strong single-hop seed; likely to support a<br/>         realistic clean question and evidence span<br/>         4 = good seed; worth authoring<br/>         3 = borderline; maybe useful but weak/noisy<br/>         2 = probably too weak or trivial<br/>         1 = bad seed; no realistic clean single-hop<br/>         question</p> <p>Think briefly, then output ONLY the single<br/>         rating digit (1-5) on the last line, nothing<br/>         else.</p> |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Every typed authoring prompt is followed by one shared block of constraints, given once here and appended verbatim to all 25 of them.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>RULES (apply to every typed single-hop prompt):</p> <ul style="list-style-type: none"> <li>- STANDALONE: name every body, person, place, year, law, product, case, or organization the question needs. Never say "this notice / ประกาศนี้ / ฉบับนี้ / เอกสารนี้ / ดังกล่าว" - the asker has not seen the Gazette document.</li> <li>- SELECTED PATTERN ONLY: write the question for the TASK above. If the document does not support that selected pattern with clean evidence, skip instead of switching to another pattern.</li> <li>- GOLD EVIDENCE: select exactly ONE `evidence_span`, copied VERBATIM from the document: exact words, spaces, punctuation, and line breaks. Use a short clean paragraph/table row when possible.</li> <li>- QUESTION-EVIDENCE FIT: the evidence span must directly address the question. It must contain the key named entity, status/event/rule, date/number/outcome needed to answer.</li> <li>- REALISTIC: avoid signer, registrar, docket-number-only, serial-number-only, Gazette</li> </ul> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

There is one typed prompt per notice class. The 24 substantive classes are the taxonomy of Appendix B.1, and unknown.md serves documents the cascade does not classify, which the default typed-only queue excludes. The prompts share a fixed skeleton and differ only in two places, namely the one-sentence statement of what the class records and the list of salient facts a question about that class should target. A bankruptcy judgment lists the debtor, the court, the case number, and the judgment date, while a decoration lists the recipient, the class of decoration, and the occasion. We reproduce three prompts spanning the range of class complexity and release the remaining 22 with the code and data.

**Bankruptcy judgment** A high-frequency formulaic class with stable field structure.

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>You write ONE natural Thai question with gold evidence for a Royal Gazette (ราชกิจจานุเบกษา) search dataset, using the single document below.</p> <p>DOCUMENT (doc_id: [[SEED_ID]]):<br/>         ""<br/>         [[SEED_TEXT]]<br/>         ""</p> <p>TASK - ask about a court judgment declaring a debtor bankrupt. Identify the debtor by NAME. Use the case number only as supporting context, not as the main thing being asked.</p> <p>Good question shapes:<br/>         - "ศาลพิพากษาให้บุคคล/บริษัท X<br/>         เป็นบุคคลล้มละลายตั้งแต่วันที่"<br/>         - "ในคดีหมายเลขแดง X<br/>         ศาลมีคำพิพากษาเกี่ยวกับลูกหนี้รายใด"</p> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

- "เจ้าหน้าที่เป็นโจทก์ยื่นฟ้องลูกหนี้รายใดจนศาลพิพากษาให้ล้มละลาย"

Evidence should include the debtor name and the bankruptcy judgment result. Include the effective date or case number only when it appears in the same clean span.

Example question shape:  
"ศาลล้มละลายกลางพิพากษาให้นายสมชาย ตัวอย่างเป็นบุคคลล้มละลายตั้งแต่วันที่"

Skip if the span is just a title line saying "คำพิพากษาให้ล้มละลาย" without debtor/result details.

Evidence should include both the erroneous text and the corrected text whenever possible. This pattern needs direct before/after evidence.

Example question shape:  
"ประกาศแก้คำผิดระบุให้แก่อธิบดีหน่วยงานจากข้อความใดเป็นข้อความใด"

Skip if the correction refers to another document but does not quote the corrected text.

## B.5.2 Multi-Hop Authoring

*Entity Extraction* emits the five fields that populate the discovery index of Appendix B.2. It runs at temperature 0 over the raw document text.

**Appointment** A person-centred class where the salient fact is a name-to-role transition.

You write ONE natural Thai question with gold evidence for a Royal Gazette (ราชกิจจานุเบกษา) search dataset, using the single document below.

DOCUMENT (doc\_id: [[SEED\_ID]]):  
""  
[[SEED\_TEXT]]  
""

TASK - ask about an appointment to an official position. Ask about the appointed person, position, replacement, or effective date.

Good question shapes:

- "บุคคล X ได้รับแต่งตั้งให้ดำรงตำแหน่งใด"
- "ผู้ใดได้รับแต่งตั้งเป็นตำแหน่ง X"
- "การแต่งตั้งบุคคล X ให้ดำรงตำแหน่ง X มีผลตั้งแต่วันที่"

Evidence should include the person name, appointed position, and effective date if asked. Prefer compact paragraphs or clean list rows. Do not select a span that only says "บุคคลดังกล่าว" without the appointed person's name.

Example question shape:  
"พลตำรวจโท เรวดี กุหลินเกสรได้รับแต่งตั้งเป็นกรรมการพิจารณาเรื่องร้องเรียนตำรวจท่านใด"

Skip bulk appointment lists when the only possible question is a random row with no useful retrieval value.

You are an entity extraction system for Thai Royal Gazette (ราชกิจจานุเบกษา) notices. The text comes from OCR, so it may contain typos or dropped characters.

Extract only what actually appears in the document. Never guess or invent. If there is nothing, use [].

- subject: what this notice does, summarized in one short sentence.

- about: \*\*the entities this notice is directly "about"\*\* - the main parties, the ones acted upon, or whatever this notice has legal effect on. E.g. the foundation being registered, the debtor placed under receivership, the temple being established, the person receiving a royal decoration or a promotion in rank. Include both person names (strip honorifics: นาย/นาง/นางสาว/ดร.) and full organization names.

- mentioned: other entities that are only mentioned in passing - not what the notice is about.

E.g. the agency issuing the notice, the signatory, the creditor bank, the court hearing the case, a political party referred to. If unsure whether something is 'about' or 'mentioned', put it in 'mentioned'.

Important: never put the same entity in both lists - pick one.

Titles that are not personal names, e.g. "นายทะเบียน", "นายอำเภอ", "รัฐมนตรี", must not be recorded as person names.

- ids: identifiers of the main parties - national ID number, case number, land title deed number, juristic person registration number.

Very important: case numbers and deed numbers are reused across courts/provinces, so they must always be prefixed with the court or province name.

Correct: "ศาลจังหวัดสระแก้ว คดีแพ่งแดง 207/2553" / "จังหวัดตรัง โฉนด 12345"

Wrong: "คดีแพ่งแดง 207/2553" / "โฉนด 12345"  
National ID numbers need no prefix - just the 13 digits.

If you cannot tell which court or province it belongs to, omit that identifier.

- refs: names of other laws/notices this document cites. Copy the full name including the B.E. year.

E.g. "พระราชบัญญัติคณะสงฆ์ พ.ศ. 2505"

Respond with JSON only.

**Correction** An amendment class that refers to an earlier notice, which is the single-hop analogue of the multi-hop correction pattern.

You write ONE natural Thai question with gold evidence for a Royal Gazette (ราชกิจจานุเบกษา) search dataset, using the single document below.

DOCUMENT (doc\_id: [[SEED\_ID]]):  
""  
[[SEED\_TEXT]]  
""

TASK - ask about an official correction notice (แก้คำผิด). Ask what text was corrected from and to.

Good question shapes:

- "ประกาศแก้คำผิดแก้ไขข้อความเดิมจากอะไรเป็นอะไร"
- "คำที่พิมพ์ผิดในประกาศเดิมเกี่ยวกับ X ถูกแก้เป็นข้อความใด"
- "ราชกิจจานุเบกษามันแก้คำผิดระบุรายการที่แก้ไขว่าอย่างไร"

Join *Relevance Screening* scores a candidate pair before authoring. The `[[RELATION]]` placeholder receives the pattern that produced the pair.

Two Royal Gazette documents are being considered as a candidate bundle for a multi-hop QA dataset. This is the final link gate before an author spends budget on them.

Claimed relation:

`[[RELATION]]`

Rate ONE thing: do BOTH documents support distinct facts about the SAME specific entity such that a realistic question genuinely needs both?

Pass only when the shared entity is specific and central to at least two notices and those notices provide distinct, coherent facts: a lifecycle, history, amendment, same case/asset, or a genuinely meaningful cross-context role. A group may contain more than two documents; it passes when it has a coherent core chain and the additional documents concern that same entity. Do not reward broad topical similarity.

Reject a person or company that is merely one row in an unrelated roster; generic professions, products, places, offices, funds, or categories; reused local registration/case numbers without unambiguous identity; issuers, signers, or cited-law mentions; template/continuation pages; and two documents that state the same fact with no multi-hop reasoning.

Rate from 1 to 5:

5 = strong: clearly the same specific entity and distinct facts need both documents  
4 = good: strongly likely despite OCR ambiguity  
3 = plausible but not proven; normally not worth authoring  
2 = probably too weak or mostly coincidental  
1 = clearly unrelated, generic, duplicate-like, or fake

Think briefly, then output ONLY the single rating digit (1-5) on the last line, nothing else.

The shared multi-hop constraints, appended to each pattern prompt.

RULES (apply to every multi-hop prompt):

- STANDALONE: name every body / person / place / year / law the question needs. Never say "this notice / ฉบับนี้ / ประกาศนี้" - the asker has not seen any Gazette document.
- QUESTION-EVIDENCE FIT: the selected evidence spans must address the question directly. They should contain the key entities, events, rules, statuses, dates, or official notices that a retriever should return for this question.
- EVIDENCE: copy each `evidence\_span` as ONE SHORT CONTIGUOUS block directly from the document - exact OCR words, Thai/Arabic digits, punctuation, spaces, and line breaks; no edits, summaries, translations, or ellipses. Do not reconstruct a list row from memory. If you cannot copy an exact block that shows the entity and its event, set `skip` to true. One clean block per selected

document.

- A roster name by itself is NOT evidence of an appointment, office, award, court, year, quota, or change. Expand the same contiguous block to include the nearest heading/column/role/event text. Likewise, an unlabelled number is not evidence of an amount or category. If the needed label and the entity cannot fit in one clean contiguous block, skip.
- MULTI-HOP: select  $\geq 2$  documents - the minimal set whose spans together answer the question.
- REAL LINK: the selected documents must have the relationship required by the task. The evidence spans may be short question-supporting excerpts and do not need to repeat the full shared context, but they must not contradict the claimed relationship.
- REALISTIC: avoid citation/date/signature trivia, enabling-authority sections, case numbers, registration numbers, roster/list order, and "which document amended which" questions. Prefer a substantive status, rule, duty, right, fee, penalty, deadline, eligibility, or consequence.
- VOICE: natural typed Thai, not legalese (avoid ดำรงตำแหน่ง, ผู้ซึ่ง). Arabic numerals are fine.
- Skip weak material: if the documents do not support a realistic, non-trivial question someone would really ask, set `skip` to true, put a short reason in `reason`, and leave `question` and `selected` empty.
- Do not write an answer. The dataset is judged by question quality and gold evidence quality.
- Before final JSON, silently check the authoring requirements: does the question name the context it needs? Does the gold evidence fit it? Is the document link real and the Thai realistic and clean? If any requirement is only borderline, fix it or skip.

Reply with ONE VALID JSON OBJECT ONLY - no prose, no code fences. Always include exactly these keys: `skip`, `reason`, `question`, `selected`. `selected` must always be an array.

Good output when keeping:

```
{
  "skip": false,
  "reason": "",
  "question": "คำถามภาษาไทย",
  "selected": [
    {
      "doc_id": "2012-000001.pdf",
      "evidence_span": "ข้อความหลักฐานแบบคัดลอกตรง"
    },
    {
      "doc_id": "2013-000002.pdf",
      "evidence_span": "ข้อความหลักฐานแบบคัดลอกตรง"
    }
  ]
}
```

Good output when skipping:

```
{
  "skip": true,
  "reason": "เอกสารไม่สัมพันธ์กันพอสำหรับคำถามหลายช่วง",
  "question": "",
  "selected": []
}
```

Evidence contrast (the wording is illustrative; always copy the actual OCR exactly):

- BAD: `๒๖. นาง ก.` for a question asking which court appointed her. The span proves only that her name occurs in a list.
- GOOD: `เรื่องแต่งตั้งผู้พิพากษาสมทบในศาล... \n... \n๒๖. นาง ก.` because the same contiguous block identifies both the event/court and the person.
- BAD: `บริษัท ก. จำกัด` for a question asking its quota category or unit.
- GOOD: include the contiguous table heading that labels the category/unit as well as that row.
- BAD: ask how a role changed when both rows state the same role, or claim year ๒๕๖๗ when the

source says ๒๕๖๓. Never turn OCR uncertainty into a factual premise; reframe truthfully or skip.

supports, so the same-person link is visible without relying on hidden metadata. Do not ask for the national ID or case number.

There is one authoring prompt per join pattern. Five patterns are typed by the relation that links the pair, covering law amendment, foundation, correction, bankruptcy case, and bankruptcy lifecycle. The sixth is the untyped ENTITY BRIDGE pattern, which additionally receives the normalized bridge string used to propose the cross-document question. We reproduce the untyped entity-bridge prompt and one typed pattern for contrast. The other four are released with the code and data.

You write ONE natural Thai question with gold evidence for a Royal Gazette (ราชกิจจานุเบกษา) search dataset, using the related documents below.

[[DOCS]]

TASK – these notices were linked because they share the same rare entity: "[[BRIDGE]]". That shared entity is the bridge. Write a question about that entity whose answer needs BOTH documents – something a journalist, auditor, lawyer, or ordinary citizen following this person / organisation / case / law would actually ask. Typically: what happened to it, and what happened to it NEXT – a status, a change, a consequence, an outcome across the two notices.

FIRST, check the link is real. The two notices were joined mechanically on a shared name or number, and OCR makes mistakes. If they are actually about DIFFERENT entities that merely share a similar name, or if the shared entity is incidental boilerplate in one of them (a signer, an enabling authority, a bank named only as a payment channel), then there is no real link – set `skip` to true and say so. Do not invent a connection to justify the pair.

Name the shared entity explicitly in the question. Each selected span must show that entity plus the event it supports, so the link is visible from the spans alone.

You write ONE natural Thai question with gold evidence for a Royal Gazette (ราชกิจจานุเบกษา) search dataset, using the related documents below.

[[DOCS]]

TASK – these notices are about the SAME person's bankruptcy (same national ID), at different stages. Ask about that person's bankruptcy STATUS over time – was the named person declared bankrupt, and what is the later status (still bankrupt, discharged, or the bankruptcy annulled)? Select evidence for the change in status and, where stated, the effective date. Identify the person by NAME, not by national ID or case number.

Each selected span must include the person's name plus the bankruptcy status/action/date it

### B.5.3 Semantic Quality Review

Each check returns a single rating from 1 through 5 on its last line, and default automated acceptance requires 4 or 5 from every applicable check. STANDALONE, EVIDENCE-FIT, and REALISTIC apply to both authoring paths; TRUE MULTI-HOP applies to multi-hop items alone.

#### Standalone

You rate ONE thing about a Thai Royal Gazette search question: is it STANDALONE?

A STANDALONE question names every body, person, place, year, or law needed to understand it WITHOUT seeing any document, and never refers to "this notice / ประกาศนี้ / ฉบับนี้ / เอกสารนี้ / ดังกล่าว".

Rate from 1 to 5:

5 = fully standalone, names everything needed  
3 = borderline, missing some context  
1 = meaningless without an unseen document, or says "this notice"

Think briefly, then output ONLY the single rating digit (1-5) on the last line, nothing else.

#### Evidence-Fit

You rate ONE thing: do the GOLD EVIDENCE spans address the QUESTION effectively?

The evidence is good when the selected spans are the right material a system should retrieve for the question: they contain the key entities/events/rules/statuses/dates needed to answer, or they are clearly the official notices a reader must use together to answer.

For single-hop questions, be strict: a short span that answers only the date/number but omits the named person/organization/product/place from the question is at best a 3. If the question asks for an amount, penalty, fee, deadline, effective date, status, or list of required items, the span must actually contain that requested fact. Related but incomplete clause fragments are at best a 3.

Apply the same strictness to EVERY span in a multi-hop question. A bare name in a roster does not prove the appointment, role, court, award, year, or status asked about. An unlabelled table number does not prove a quota, amount, unit, or category. The selected contiguous span itself must carry the nearby heading, column label, role, or event context needed to interpret it; do not silently borrow that context from an unseen document title or text outside the gold span. If one required hop has only a name/number while the other hop is good, score at most 3.

Examples:

- Question asks whether a named person was appointed to the same court twice; both spans contain only numbered name rows -> 2 or 3, not 5.
- Question compares two offices and each span says 'Name - Office' -> 5.
- Question asks quota category/tonnage but rows contain only company plus an unlabelled decimal -> at most 3.

Rate from 1 to 5:

- 5 = excellent gold evidence; the spans directly and sufficiently address the question
- 4 = good evidence; mostly sufficient, with only minor missing context or wording friction
- 3 = partially useful; related to the question but missing an important fact, date, entity, or step
- 2 = weak; only loosely related or one span is useful while the other is mostly irrelevant
- 1 = wrong evidence; does not answer/address the question, or points to the wrong entity/case/notice

Penalize only if the question and selected evidence themselves do not line up.

Think briefly, then output ONLY the single rating digit (1-5) on the last line, nothing else.

**Realistic** This check receives the source excerpts alongside the question, because a question can read as natural while resting on a role or a year that OCR noise invented, and only the source reveals that.

You rate ONE thing about a Thai Royal Gazette question: is it a REALISTIC, non-trivia question written in clean Thai?

Rate from 1 to 5:

- 5 = a real person (lawyer, journalist, official, business owner, citizen) would genuinely ask it, and the question is clean readable Thai
- 3 = somewhat niche or slightly awkward, but plausible
- 1 = signer / registrar / case-number / serial / roster / "who is first in the list" trivia, OR the question contains OCR garble or nonsense, falsely calls an award class a job/status, presupposes a change when it actually asks whether an unchanged row changed, or asserts a year/event that conflicts with the supplied Gazette material

Do not reward grammatical fluency when the underlying proposition is nonsensical or visibly manufactured from OCR. A clean, substantive comparison of two real roles, awards, rules, or events is realistic; a comparison built from corrupted entity/type names or invented chronology is not.

Think briefly, then output ONLY the single rating digit (1-5) on the last line, nothing else.

**True Multi-Hop** The `[[RELATION]]` placeholder receives the join pattern, and the check reads the full selected documents when scoring whether every cited document is necessary.

Two Royal Gazette documents were combined to write ONE multi-hop question. Their claimed relation:

`[[RELATION]]`

You rate ONE thing: is the question GENUINELY multi-hop - does answering it actually require BOTH documents, not just one?

A question is genuinely multi-hop when a necessary part of the answer comes from EACH document:  
a fact, entity, status, date, amount, or step in one document combined with a different fact in the other. Neither document alone is enough.

BOTH of these count as genuinely multi-hop - do not penalize either:

- COMPOSITION (bridge): a fact in one document (a name, case no., registration, pointer) is what lets you locate or interpret the answer in the other;
  - COMPARISON / TWO-EVENT: the question compares two versions, or asks about two distinct events or states, where EACH document supplies a necessary part of the answer and neither alone is enough.
- A retriever must find both. This is legitimate multi-hop even though no bridge is traversed.

It is NOT genuinely multi-hop when:

- one document already contains everything needed to answer, and the other is decorative;
- the second document only confirms or repeats a fact the first already states;
- the question only NARRATES a fact from one document as background while the whole answer sits in the other;
- the question's PREMISE IS FALSE: the two documents are not actually related the way the question asserts (different people, different cases, different provisions), so the comparison or link it asks about does not exist. Score these 1 - a false premise is worse than a single-hop question.

The test is simply: can ONE document alone fully answer this? If yes -> 1. If a necessary part of the answer is only in the other document -> it is multi-hop, whether by bridge or by comparison.

Use the full documents (not only the short selected quotes) to judge whether both are truly load-bearing. Do not penalize a document just because its contribution is a short row, name, date, or amount - a small but necessary fact still counts as a required hop.

Do NOT grade evidence-span quality here. If the full documents really contain two distinct facts needed by the question, score the multi-hop structure highly even when a selected quote is too short; the separate evidence-fit judge owns that failure. Conversely, a chronological preamble from document A does not make a question multi-hop when the requested answer is wholly in B.

Rate from 1 to 5:

- 5 = clearly requires both documents; each contributes a necessary, distinct part of the answer
- 3 = weak; one document does most of the work, the other adds only minor context
- 1 = single-hop in disguise: one document alone answers it, or the two halves are unrelated

Think briefly, then output ONLY the single rating digit (1-5) on the last line, nothing else.

## C Units for Evidence Retrieval

**Word chunks** The Thai `newmm` tokenizer identifies non-whitespace words while preserving their character offsets. Consecutive windows contain the requested number of words. Overlapping variants advance by the chunk size minus the overlap, and a final window containing only repeated overlap is omitted.

**Sentence chunks** The Thai `crfcut` sentence tokenizer segments the document, while `newmm` counts words for the packing budget. Consecutive sentences are packed into a chunk until the next sentence would push the chunk past the requested word count. The requested size is therefore an upper bound rather than an exact length, and boundaries always fall at a sentence end. A sentence longer than the requested size becomes its own chunk and is not split further, which is what produces the small number of oversized units on documents where the tokenizer finds no boundary, such as tabular filings and dotted form templates. At a 200-word setting this yields 55,465 units against 53,196 for the word splitter, so the two conditions differ in boundary placement rather than in unit count. Overlap for this splitter is counted in sentences rather than words, so it is not comparable to the word overlap ablation, and we hold it at zero.

**Contextual chunks** Each base chunk is paired with the complete source document. DeepSeek V4 Flash generates a short context that is prepended to the chunk for embedding and lexical indexing only. The original chunk remains the evaluation text, supplies the source offsets, and is the text passed to the Evidence-Fit judge and to the downstream generator, so no contextual prefix is ever scored or read at answer time. Batch generation falls back to individual requests when a batch response fails.

**Atomic propositions** DeepSeek V4 Flash emits self-contained factual statements paired with source excerpts. The system retrieves each statement, while evaluation uses the located excerpt and its offsets. Exact excerpt matching is attempted first, followed by fuzzy alignment. Of 201,740 extracted units, 200,903 initially received a non-empty interval. The initial alignment failure rate was therefore 0.415% (837 units). We performed an agent-

assisted manual audit of every failure against its source document. Deterministic exact and fuzzy searches proposed candidate windows, but a reviewer accepted a repair only when one contiguous passage supported the complete proposition. A mechanical gate then confirmed that the saved quote equalled the half-open source slice. We independently re-reviewed 104 accepted spans that exceeded 1,500 characters or five times the failed excerpt length. The manual repair recovered 411 intervals, classified 412 units as ambiguous because their evidence was non-contiguous or structurally reordered, and found 14 unsupported units. This leaves 201,314 units with a verified interval, a final interval rate of 99.789%, and a residual rate of 0.211%. A unit without a located interval contributes no overlap, so it scores as retrieved text that recovers no evidence. The immutable original artifact, repaired copy, and decision-level audit are retained separately.

**Whole documents** Whole-document conditions use one unit spanning the complete source text. The unit retains the source-document identifier and covers offsets from the first through the final character. This construction uses no model.

### C.1 Retrieval-Unit Prompts

The contextual and atomic-proposition units are the only retrieval units an LLM constructs; both run DeepSeek V4 Flash. Word, sentence, and whole-document units are tokenizer-based and use no model. Both LLM prompts instruct the model to answer in Thai when the source is Thai, so the generated text matches the language of the questions.

**Contextual chunks** The single-chunk prompt.

```
<document>
{document}
</document>

Here is the chunk we want to situate within the
whole document:

<chunk>
{chunk}
</chunk>

Please give a short succinct context to situate
this chunk within the overall document for the
purposes of improving search retrieval of the
chunk. Write the context in the same language
as the chunk. For Thai Royal Gazette text,
answer in Thai. Answer only with the succinct
context and nothing else.
```

The batch prompt, used by default and falling back to the single-chunk prompt when a batch response fails to parse.

```
<document>
{document}
</document>
```

Below are {n} numbered chunks taken from the document above.

```
<chunks>
{numbered_chunks}
</chunks>
```

For EACH chunk, give a short succinct context to situate it within the overall document, for the purpose of improving search retrieval of that chunk. Write each context in the same language as the chunk. For Thai Royal Gazette text, answer in Thai.

Answer with exactly {n} lines, one per chunk, each formatted exactly as:  
<number>. <context>

Do not repeat the chunk text. Do not include anything else before or after the {n} lines.

## Atomic propositions The default extraction prompt.

You are a legal analyst specializing in Thai Royal Gazette documents. Your task is to extract discrete atomic propositions from the text below. An atomic proposition is a self-contained factual statement.

**\*\*Extraction rule:\*\***  
Extract a proposition from every sentence or clause that states something concrete. When in doubt, extract it. Do not filter at this stage – the requirements below handle quality.

**\*\*Coverage requirement:\*\***  
You MUST extract propositions from the entire document. Mentally divide the document into sections (header, rationale, legal authority, definitions, operative clauses, conditions, signatories) and verify at least one proposition has been extracted from every section containing substantive information before finalizing output. If any section produced zero propositions, return to it and extract what you missed.

**\*\*Requirements for each proposition:\*\***  
– Express exactly one specific claim (one subject, one predicate)  
– Use full names, titles, and official designations – never pronouns or vague references  
– Include all dates, numbers, legal references, and qualifiers exactly as written  
– If multiple items belong to the same named category or list, group them into one proposition  
– Do not extract the same proposition twice – if a proposition has already been extracted, skip the duplicate  
– Maintain the original language of the source text

**\*\*Self-containment rule – MUST pass before including any proposition:\*\***  
Apply this test to every proposition before writing it: **"Could someone who has never seen this document understand this proposition completely, with no missing who, what, when, where, or under what authority?"**  
If the answer is no – add the missing elements. A proposition that names a person MUST also include their affiliation and the announcement context. A proposition that states an award MUST also include the year, the awarding authority, and the recipient's full affiliation. Never strip document-level context (announcement title, year, issuing body) from any proposition.

**\*\*Excerpt verbatim rule – MUST pass before writing any excerpt:\*\***  
The `excerpt` field is a byte-for-byte copy of the source text. Apply this test before writing every excerpt: **"Is every character, space, line break, and punctuation mark in my excerpt identical to the source?"** If no – rewrite it.

NEVER do any of the following to the excerpt:  
– Paraphrase, summarize, or reword any part  
– Omit any words, characters, or symbols  
– Shorten or truncate for brevity  
– Normalize whitespace, indentation, or line breaks  
– Correct spelling, punctuation, or formatting  
– Replace or remove special characters

If the source text contains irregular spacing, mid-word line breaks, unusual indentation, or Thai numeral formatting – reproduce all of it exactly. The excerpt must be indistinguishable from a direct paste of the source.

**\*\*Format lock – output this exact JSON structure. No markdown, no explanation, no text outside the JSON:\*\***

```
```json
{
  "propositions": [
    {
      "proposition": "Self-contained factual statement in Thai",
      "excerpt": "Exact byte-for-byte copy of the source text – no changes of any kind"
    }
  ]
}
```

**\*\*Grounding rule:\*\*** Extract only propositions explicitly stated in the text. Do not infer, extrapolate, or add context from outside the provided text.

**\*\*Text to process:\*\***  
{input}

## A variant selected for DeepSeek, which follows a terser specification more reliably.

You extract retrieval-ready atomic propositions from Thai Royal Gazette text. An atomic proposition is a self-contained factual statement.

Return only one JSON object with this shape:

```
{
  "propositions": [
    {
      "proposition": "A self-contained factual statement in the source language",
      "excerpt": "A short exact evidence span copied from the source"
    }
  ]
}
```

Rules for good propositions:

- Extract concrete legal/administrative facts that would help answer search queries.
- Each proposition must be self-contained: include the announcement/order name, issuing body, names, roles, dates, numbers, legal references, and conditions needed to understand it.
- Keep one proposition to one main claim. If a table row describes one entity, make one proposition for that row.
- If one source sentence or paragraph lists several people/items under the same action, group them into one proposition instead of making separate propositions that require skipped evidence.
- Preserve names, dates, numbers, legal references, and OCR spelling exactly as shown in the source. Do not correct the source.

- Do not invent missing context, normalize dates, translate text, or add outside knowledge.
- Prefer 3-12 high-value propositions for a short notice; use up to 20 only when the text contains a table/list with many distinct rows.
- Skip pure page furniture unless it identifies the notice, issuing body, date, case number, or legal authority.

Rules for `excerpt`:

- The excerpt is evidence for locating the proposition in the original text.
- Copy a short continuous span from the source text exactly.
- Choose the smallest span that supports the proposition, usually one sentence, clause, heading, or table row.
- The `excerpt` value must be a literal substring of the source text.
- Preserve source line breaks inside `excerpt`.
- If a heading is split across several lines, copy those line breaks or choose one exact line as the excerpt.
- Never use ellipses (`...`), bracketed omissions, summaries, markdown fences, or paraphrases in `excerpt`.
- If no single continuous source span supports every detail in a proposed proposition, either make the proposition less detailed or group the source list into one proposition with one continuous excerpt.

Output rules:

- Output valid JSON only.
- No markdown.
- No explanation.
- No trailing text.

Text:

{input}

## D Evidence-Retrieval Implementation

**Libraries** Sentence segmentation uses `crfcut`, while word segmentation and BM25 tokenization use `newmm`, both from PyThaiNLP (Phatthiyaphai-bun et al., 2023). Dense retrieval uses an exact FAISS inner-product index (Douze et al., 2026), so the reported dense numbers reflect the encoder rather than an approximate-search error. BM25 uses the `bm25s` Lucene scoring variant with  $k_1 = 1.0$  and  $b = 0.9$ , held fixed across all conditions. We checked this choice with a 60-point grid search over  $k_1 \in [0.2, 1.8]$  and  $b \in [0.3, 1.0]$ , scoring BM25-only hit rate at  $k = 4$  over 200-word units on a 500-question development split (seed 42). The best development setting,  $k_1 = 0.3$  and  $b = 0.4$ , gains 2.0 points there but only 0.5 points on the 1,500 held-out questions (0.825 against 0.820), and the full grid spans 0.766 to 0.816. BM25 performance is therefore insensitive to both parameters on this corpus, and we report the fixed values rather than re-tuning. Multivector retrieval uses the PLAID (Santhanam et al., 2022) index provided by PyLate (Chaffin and Sourty, 2025), which exposes late-interaction indexing over BGE-M3 token vectors without a ColBERT-specific reimplementa-

**Models** Table 4 lists the checkpoints used in evidence retrieval and downstream evaluation.

| Role                | Checkpoint              |
|---------------------|-------------------------|
| Default encoder     | BAAI/bge-m3             |
| Multivector encoder | BAAI/bge-m3             |
| Encoder ablation    | Qwen/Qwen3-Embedding-4B |
| Encoder ablation    | voyage-4-nano           |
| Reranker            | Qwen/Qwen3-Reranker-4B  |
| Answer generator    | google/gemma-4-31B-it   |
| Pairwise judge      | claude-opus-5           |

Table 4: Models by role. BGE-M3 supplies both pooled and token vectors.

### D.1 Reranking Prompt Components

The reranked conditions of Section 4.3 score each question and unit pair with the Qwen3 reranker, whose scaffolding is fixed across all conditions. The system prompt.

```
<|im_start|>system
Judge whether the Document meets the
requirements based on the Query and the
Instruct provided. Note that the answer can
only be "yes" or "no".<|im_end|>
<|im_start|>user
```

The user prompt, which carries the task instruction, the question, and the candidate unit.

```
<Instruct>: {instruction}
<Query>: {query}
<Document>: {document}
```

The task instruction substituted into the user prompt.

```
Given a web search query, retrieve relevant
passages that answer the query
```

The suffix that elicits the yes-or-no token whose probability becomes the relevance score.

```
<|im_end|>
<|im_start|>assistant
<think>

</think>
```

## E Additional Results and Analyses

### E.1 Delivered Text and Efficiency

Table 5 reports the text each configuration actually delivers. Query time is milliseconds per question, measured over the same 2,000 dataset questions used for evidence-retrieval scoring. It includes query encoding, candidate retrieval, fusion, and reranking when applicable. LLM construction of

contextual and atomic-proposition units is excluded as offline preprocessing. All configurations ran on one workstation with an NVIDIA RTX 3090 and an AMD Ryzen 5 5600X. These are single-run wall-clock measurements on shared hardware, so they indicate relative order of magnitude rather than rigorous latency measurements, and we round them to two significant figures. We omit index construction time because indices are cached and shared across configurations, which makes the recorded build times reflect cache state rather than the cost of the method.

| Configuration                                                | Words/q |
|--------------------------------------------------------------|---------|
| <i>Word chunks, single-vector (bge-m3)</i>                   |         |
| 50w ( $k=16$ )                                               | 767     |
| 100w ( $k=8$ )                                               | 726     |
| 200w ( $k=4$ )                                               | 669     |
| 400w ( $k=2$ )                                               | 484     |
| <i>Word 200w, one-change ablations</i>                       |         |
| + overlap-50                                                 | 667     |
| + sentence boundaries                                        | 628     |
| voyage-4-nano embed                                          | 659     |
| qwen3-4b embed                                               | 653     |
| multivector (ColBERT)                                        | 674     |
| <i>Word 200w, sparse-dense fusion</i>                        |         |
| hybrid                                                       | 683     |
| hybrid + qwen rerank                                         | 674     |
| <i>Contextual chunks (200w, bge-m3)</i>                      |         |
| vector                                                       | 661     |
| hybrid                                                       | 650     |
| hybrid + qwen rerank                                         | 660     |
| <i>Atomic-proposition chunks (bge-m3), <math>k=16</math></i> |         |
| vector                                                       | 494     |
| hybrid                                                       | 472     |
| hybrid + qwen rerank                                         | 459     |
| <i>Whole-document chunks (bge-m3), <math>k=4</math></i>      |         |
| dense                                                        | 1,233   |
| multivector (ColBERT)                                        | 1,209   |

Table 5: Delivered text per configuration, the mean number of words returned per question over a 200-question sample. The 800-word budget is nominal, because units at document ends are short, so no two configurations deliver the same amount of text and both span metrics move with what is delivered.

## E.2 Paired Uncertainty Analysis

Table 8 reports each one-change contrast as a comparison-minus-baseline difference with a 95% interval. The question is the sampling unit, configurations stay paired by question identifier, and the per-question difference is formed before resampling. Intervals are percentile intervals over 20,000 bootstrap replicates drawn within the single-hop and multi-hop strata at their observed sizes (1,500 and 500 for the span metrics, 75 and 25 for the

| Configuration                                                | Query (ms/q) |
|--------------------------------------------------------------|--------------|
| <i>Word chunks, single-vector (bge-m3)</i>                   |              |
| 50w ( $k=16$ )                                               | 4.8          |
| 100w ( $k=8$ )                                               | 3.4          |
| 200w ( $k=4$ )                                               | 2.8          |
| 400w ( $k=2$ )                                               | 2.5          |
| <i>Word 200w, one-change ablations</i>                       |              |
| + overlap-50                                                 | 2.9          |
| + sentence boundaries                                        | 3.1          |
| voyage-4-nano embed                                          | 4.6          |
| qwen3-4b embed                                               | 18           |
| multivector (ColBERT)                                        | 270          |
| <i>Word 200w, sparse-dense fusion</i>                        |              |
| hybrid                                                       | 3.7          |
| hybrid + qwen rerank                                         | 2,000        |
| <i>Contextual chunks (200w, bge-m3)</i>                      |              |
| vector                                                       | 2.8          |
| hybrid                                                       | 3.5          |
| hybrid + qwen rerank                                         | 2,100        |
| <i>Atomic-proposition chunks (bge-m3), <math>k=16</math></i> |              |
| vector                                                       | 5.6          |
| hybrid                                                       | 7.1          |
| hybrid + qwen rerank                                         | 3,100        |
| <i>Whole-document chunks (bge-m3), <math>k=4</math></i>      |              |
| dense                                                        | 2.5          |
| multivector (ColBERT)                                        | 360          |

Table 6: Query cost per configuration. Lower is better. Cross-encoder reranking dominates query time, costing roughly two to three orders of magnitude more than the fusion step it reranks, and late-interaction multivector retrieval sits between the two.

judge), so every replicate preserves the scenario mix of the evaluation set. One shared resample is applied across all contrasts within a replicate.

We report intervals rather than hypothesis tests, and we apply no correction for the number of contrasts, so an interval that excludes zero is one marginal comparison rather than a family-wise claim. The judge intervals rest on 100 ratings and resolve ten of the eighteen contrasts, which is why the judge score is treated as descriptive throughout. Comparing two contrasts by whether their intervals overlap is not a test of their difference, and we make no such comparison in the text.

## E.3 Downstream Answer Evaluation

The generator uses temperature zero, a 512-token limit, and no abstention or citation instruction. Its prompt contains only the question and retrieved units. The judge sees gold evidence for comparison but not retrieved units or system names. Answer order and the round-robin opponent for each question are deterministically balanced with seed 42.

The run contains 1,800 comparisons, of which 62.2% are ties. Of these, 563 pairs are byte-

identical answers recorded as ties without a judge call, leaving 1,237 comparisons judged by Claude Opus 5. Against contextual hybrid with Qwen, contextual hybrid retrieval records 1 win, 1 loss, and 8 ties, while document multivector records 1 win, 1 loss, and 9 ties. Ties dominate every head-to-head record among the leading systems.

#### **E.4 Performance by Notice Type**

Table 9 compares baseline (ordinary 200-word dense retrieval) vs. recommended (contextual 200-word hybrid retrieval) by notice type. We classify gold source documents using the existing rules, resolving multiple matches by routing order. Questions whose sources share a primary type receive that label; differing types form the mixed group, and any unclassified source places a question in the unknown group.

Recall improves in 21 of 24 named types and is unchanged in three. Among types with at least 30 questions, incapacity orders (+29.42 pp) and association registrations (+28.83 pp) show the largest recall gains. Precision gains are less uniform: bank-branch notices lose 1.36 pp and condominium notices lose 0.38 pp despite improved recall.

Overall changes are +11.89 pp in recall, +0.91 pp in precision, and +0.73 Evidence-Fit points. Sparse ratings, automatic labels, and differences in question composition limit per-type interpretation.

| Configuration                            | Query (ms/q) |
|------------------------------------------|--------------|
| Word 50w ( $k=16$ ) dense                | 4.8          |
| Word 100w ( $k=8$ ) dense                | 3.4          |
| Word 200w ( $k=4$ ) dense                | 2.8          |
| Word 400w ( $k=2$ ) dense                | 2.5          |
| Word 200w overlap-50                     | 2.9          |
| Word 200w sentence boundaries            | 3.1          |
| Word 200w voyage-4-nano embed            | 4.6          |
| Word 200w qwen3-4b embed                 | 18           |
| Word 200w multivector (ColBERT)          | 270          |
| Word 200w hybrid                         | 3.7          |
| Word 200w hybrid + qwen rerank           | 2,000        |
| Contextual 200w vector                   | 2.8          |
| Contextual 200w hybrid                   | 3.5          |
| Contextual 200w hybrid + qwen rerank     | 2,100        |
| Atomic propositions vector               | 5.6          |
| Atomic propositions hybrid               | 7.1          |
| Atomic propositions hybrid + qwen rerank | 3,100        |
| Whole document dense                     | 2.5          |
| Whole document multivector (ColBERT)     | 360          |

Table 7: Query cost per configuration. Lower is better. Word and contextual chunks use BGE-M3 unless an alternative encoder is named. Atomic propositions use  $k = 16$ ; 200-word and whole-document configurations use  $k = 4$ . Cross-encoder reranking dominates query time.

| Contrast                        | SpanRec                  | SpanPrec                 | EvFit                 |
|---------------------------------|--------------------------|--------------------------|-----------------------|
| <i>Retrieval unit</i>           |                          |                          |                       |
| Word 50w – word 200w            | −0.175 [−0.194, −0.157]* | −0.022 [−0.024, −0.020]* | −0.35 [−0.68, −0.02]* |
| Word 100w – word 200w           | −0.087 [−0.103, −0.071]* | −0.012 [−0.013, −0.010]* | −0.21 [−0.51, 0.10]   |
| Word 400w – word 200w           | +0.020 [0.006, 0.034]*   | +0.031 [0.029, 0.034]*   | +0.32 [0.04, 0.60]*   |
| Overlap-50 – word 200w          | +0.021 [0.012, 0.031]*   | +0.003 [0.002, 0.004]*   | +0.34 [0.14, 0.56]*   |
| Sentence bounds – word 200w     | −0.013 [−0.023, −0.003]* | +0.001 [−0.000, 0.002]   | +0.06 [−0.17, 0.29]   |
| Contextual – word 200w          | +0.059 [0.045, 0.072]*   | +0.005 [0.004, 0.007]*   | +0.34 [0.09, 0.60]*   |
| Atomic propositions – word 200w | −0.118 [−0.139, −0.096]* | +0.011 [0.008, 0.014]*   | +0.31 [−0.03, 0.66]   |
| Whole document – word 200w      | +0.099 [0.085, 0.112]*   | −0.014 [−0.015, −0.012]* | +0.50 [0.23, 0.78]*   |
| <i>Retrieval method</i>         |                          |                          |                       |
| Word hybrid – dense             | +0.054 [0.038, 0.069]*   | +0.002 [0.001, 0.004]*   | +0.36 [0.10, 0.63]*   |
| Word rerank – hybrid            | +0.052 [0.042, 0.063]*   | +0.003 [0.002, 0.004]*   | +0.03 [−0.22, 0.29]   |
| Word multivector – dense        | +0.036 [0.024, 0.047]*   | +0.001 [−0.001, 0.002]   | +0.08 [−0.13, 0.29]   |
| Contextual hybrid – dense       | +0.060 [0.046, 0.075]*   | +0.004 [0.002, 0.005]*   | +0.39 [0.17, 0.63]*   |
| Contextual rerank – hybrid      | +0.041 [0.031, 0.052]*   | +0.001 [−0.000, 0.002]   | −0.04 [−0.21, 0.12]   |
| Propositions hybrid – dense     | −0.014 [−0.025, −0.003]* | −0.000 [−0.002, 0.001]   | −0.30 [−0.52, −0.10]* |
| Propositions rerank – hybrid    | +0.061 [0.050, 0.072]*   | +0.005 [0.003, 0.007]*   | +0.22 [0.03, 0.43]*   |
| Document multivector – dense    | +0.038 [0.027, 0.049]*   | +0.002 [0.001, 0.003]*   | +0.07 [−0.06, 0.21]   |
| <i>Encoder</i>                  |                          |                          |                       |
| Voyage-4-Nano – BGE-M3          | +0.040 [0.028, 0.053]*   | +0.004 [0.002, 0.005]*   | +0.33 [0.07, 0.59]*   |
| Qwen3-4B – BGE-M3               | −0.009 [−0.023, 0.005]   | −0.001 [−0.002, 0.001]   | −0.05 [−0.33, 0.23]   |

Table 8: Paired comparison-minus-baseline differences with scenario-stratified 95% bootstrap intervals. \*marks an interval that excludes zero, uncorrected for the 54 comparisons shown.

| Notice type                     | Questions | $\Delta$ Recall<br>(pp) | $\Delta$ Prec<br>(pp) | $\Delta$ EvFit<br>(n) |
|---------------------------------|-----------|-------------------------|-----------------------|-----------------------|
| Overall                         | 2000      | +11.89                  | +0.91                 | +0.73 (100)           |
| Rule or notification            | 400       | +5.91                   | +0.57                 | +0.74 (27)            |
| Foundation registration         | 240       | +9.53                   | +1.33                 | +0.17 (12)            |
| Appointment                     | 229       | +14.76                  | +0.91                 | +0.73 (11)            |
| Local ordinance                 | 119       | +9.43                   | +0.96                 | 0.00 (5)              |
| Condominium registration        | 96        | +3.77                   | −0.38                 | 0.00 (4)              |
| Association registration        | 77        | +28.83                  | +1.42                 | +2.40 (5)             |
| Incapacity order                | 74        | +29.42                  | +1.80                 | +2.00 (2)             |
| Association / foundation change | 71        | +9.90                   | +0.52                 | +1.33 (3)             |
| Bankruptcy discharge            | 66        | +2.18                   | +0.18                 | 0.00 (1)              |
| Company registration list       | 62        | +23.93                  | +1.97                 | 0.00 (2)              |
| Land title notice               | 49        | +16.33                  | +0.51                 | 0.00 (2)              |
| Bankruptcy judgment             | 46        | +21.74                  | +1.18                 | 0.00 (1)              |
| Bank branch                     | 40        | +5.79                   | −1.36                 | +0.50 (2)             |
| Industrial standard             | 38        | +13.16                  | +0.39                 | +1.33 (3)             |
| Temple establishment            | 35        | 0.00                    | +0.07                 | 0.00 (4)              |
| Registration revocation         | 28        | +16.46                  | +2.25                 | 0.00 (2)              |
| Decoration                      | 25        | +14.29                  | +1.32                 | +4.00 (1)             |
| Missing person order            | 25        | +16.00                  | +0.59                 | NA (0)                |
| Bankruptcy receivership         | 25        | +32.00                  | +1.44                 | NA (0)                |
| Mining concession               | 11        | +18.18                  | +0.78                 | NA (0)                |
| Labor union                     | 9         | +9.92                   | −0.14                 | NA (0)                |
| Catholic church recognition     | 9         | 0.00                    | −0.13                 | NA (0)                |
| Election boundary               | 4         | +25.00                  | +0.83                 | NA (0)                |
| Correction                      | 1         | 0.00                    | +0.11                 | NA (0)                |
| Mixed notice types              | 90        | +8.67                   | +1.52                 | +1.43 (7)             |
| Unknown / incomplete            | 131       | +13.87                  | +1.77                 | +0.67 (6)             |

Table 9: Baseline vs. recommended retrieval ( $k = 4$ ). Deltas are recommended minus baseline: percentage points for recall and precision, rating points for Evidence-Fit (1–5). Questions gives the paired sample for recall and precision. Parentheses give the number of questions rated for Evidence-Fit; NA (0) indicates no ratings. Changes use unrounded scores. Judge sampling was stratified by scenario.