

# Deka Retrieval: Component-Level Criteria for Mining Legal Case Pairs in Thai Supreme Court Decisions

Pasin Buakhaw<sup>1</sup> Pawitsapak Akarajardwong<sup>2</sup> Chompakorn Chaksangchaichot<sup>2</sup>

<sup>1</sup>Computer Engineering and Digital Technology, Chulalongkorn University, Bangkok, Thailand

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

6833184221@student.chula.ac.th

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

## Abstract

Legal case retrieval (LCR) defines relevance beyond text similarity, judgments may describe near-identical events yet be decided under different charges. Existing synthesis pipelines mine training pairs from text similarity and shared charges, which separates positives from negatives but leaves the difficulty of each pair implicit. We present **Deka Retrieval**, an LCR benchmark on Thai Supreme Court decisions with an LLM relevance cases mining pipeline for a civil-law jurisdiction. We decompose each judgment into case components (Charge, Fact, Reasoning, Decision) and split easy/hard level on exist binary signals and introduce an *anomaly* variant a conflict judgment that shares the anchor's charge and facts but reaching a different reasoning and decision. Retriever performance follows query length and legal terminology: Fact > Lawyer expert > Lay people. Fine-tuning gains most on the weakest type, lifting Lay R@10 by  $2.6 \times$  ( $7.6 \rightarrow 19.5$ ) versus  $1.5 \times$  on Lawyer and  $1.3 \times$  on Fact. Training on the shortest query also transfers to the longer ones, but not the reverse. Adding *anomaly* cases in training as positive degrades the overall performance, compare with without it. Our dataset and models are available at <https://huggingface.co/collections/VISAI-AI/deka-retrieval>

## 1 Introduction

Legal case retrieval (LCR) is a fundamental task in legal research, allowing practitioners to identify prior judgments related to the matter under consideration. Although this is a document retrieval problem, LCR has its unique challenges; case documents are generally long, packed with specific domain terminology, and governed by a notion of

relevance that extends beyond lexical or semantic similarity (Shao et al., 2023). While a statute defines the law in abstract way, judgment applies that abstract rule to what actually happened, recording the case facts, court's reasoning and outcome decision (Mandal et al., 2017). The relevance between judgments is therefore determined by legal congruity, a shared charge, provision, or rationale, so that cases with closely similar facts may nevertheless be decided differently.

These characteristics make the construction of training pairs for LCR particularly challenging. Existing synthesis pipelines mine pairs mainly from text similarity and shared charges (Deng et al., 2024b; Gao et al., 2024), this coarsely separates positives from negatives but does not guarantee legal relevance. In the general domain, grading pairs by difficulty is known to matter for dense-retrieval training (Xiong et al., 2020; Zhan et al., 2021) legal pipelines have so far left this grading implicit. In particular, semantically similar cases may be legally incongruent when their outcomes turn on different legal elements. Conversely, cases sharing the same adjudicated offense may remain relevant despite substantial differences in their surface facts. For example, under Thai criminal law, theft<sup>1</sup> and misappropriation<sup>2</sup> (embezzlement) are codified as distinct offenses that turn on whether the property was already in the offender's lawful possession, so a case's charge set precisely encodes the mode of the taking despite the scenario being semantically similar.

We therefore define explicit criteria for what counts as an easy or hard positive and negative, stated over case components rather than over surface similarity. Table 1 illustrates them on a set of embezzlement cases of an entrusted vehicle, both

<sup>1</sup><https://library.siam-legal.com/thai-law/criminal-code-theft-sections-334-336-2/>

<sup>2</sup><https://library.siam-legal.com/thai-law/criminal-code-misappropriation-sections-352-356/>

This work was done as a part of CEDT Internship Program at VISAI.AI

|                                                              |                                                                                                                                                                                                                                          |
|--------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Anchor — Embezzlement</b>                                 |                                                                                                                                                                                                                                          |
| Fact                                                         | The defendant hire-purchased a car from the owner and later dishonestly took the vehicle.                                                                                                                                                |
| Reasoning                                                    | The charge is <b>embezzlement</b> , together with a civil claim for the return of the car; the court held that the civil claim was connected to the criminal case.                                                                       |
| Decision                                                     | Judgment <b>modified</b> : the lower courts’ rulings on the civil claim were set aside and the court fees refunded; the <b>conviction</b> otherwise stood.                                                                               |
| <b>Easy Positive — Embezzlement</b>                          |                                                                                                                                                                                                                                          |
| Fact                                                         | The defendant collected a road grader that its owner had left in the victim’s custody at a gas station and kept it for himself.                                                                                                          |
| Reasoning                                                    | Although charged as theft, the court re-characterised the act as <b>embezzlement</b> , since the victim, who had lawful possession, had asked the defendant to remove the machine.                                                       |
| Decision                                                     | Judgment <b>modified</b> : the defendant was found <b>guilty</b> of embezzlement and sentenced to one year of imprisonment.                                                                                                              |
| <b>Hard Positive — Embezzlement</b>                          |                                                                                                                                                                                                                                          |
| Fact                                                         | The defendant dishonestly took the victim’s money for himself between August 2013 and May 2014.                                                                                                                                          |
| Reasoning                                                    | The charge is <b>embezzlement</b> with a connected civil claim; the court addressed a single judge’s authority over the civil claim and the interest payable on the tort debt.                                                           |
| Decision                                                     | Judgment <b>modified</b> : the defendant must pay 736,289 baht plus 7.5% annual interest, with an earlier payment of 250,000 baht credited first.                                                                                        |
| <b>Hard Negative — Theft &amp; Receiving Stolen Property</b> |                                                                                                                                                                                                                                          |
| Fact                                                         | The defendants jointly stole a car at night using another vehicle; the stolen car was subject to a hire-purchase agreement between the claimant-owner and the lessee.                                                                    |
| Reasoning                                                    | The dispute concerned the return of the forfeited car in a case involving <b>theft by vehicle</b> and <b>receiving stolen property</b> ; recovery under Section 36 was available to the owner.                                           |
| Decision                                                     | Judgment <b>reversed</b> : the seized car was <b>returned to the claimant</b> .                                                                                                                                                          |
| <b>Anomaly — Embezzlement</b>                                |                                                                                                                                                                                                                                          |
| Fact                                                         | The defendant was entrusted with possession of the victim’s rice harvester and sold it for personal gain.                                                                                                                                |
| Reasoning                                                    | On the <b>embezzlement</b> charge, the court found that the harvester proved to have been sold was a <i>different</i> machine from the one identified in the indictment; the proven facts therefore materially departed from the charge. |
| Decision                                                     | Judgment <b>affirmed</b> : the <b>acquittal</b> stood.                                                                                                                                                                                   |

Table 1: Heuristic criteria and anomaly examples on embezzlement cases.

the Easy and Hard Positives share the Embezzlement charge but Easy case is factually closer to the anchor (also involving a vehicle), while the Hard case involves money. Hard Negative shares the

topic a vehicle taken yet overlaps in no charge, being decided as Theft and Receiving stolen property instead.

Query type can change retrieval results (Penha et al., 2022; Sciavolino et al., 2021), and in the legal domain, as in the clinical one (Tamine and Goeuriot, 2021), a meaningful context retrieves better but may not reflect a real-world search: the symmetric setting retrieves a document from a document (Ma et al., 2021), while recent work instead poses queries by the user’s legal knowledge (Zhao et al., 2026; Paul et al., 2026; Choi et al., 2026b). In our experiment the lengthy query performs best, but the gap left by layperson keywords is the one fine-tuning can close, lifting Lay R@10 by  $2.6\times$  against  $1.5\times$  on Lawyer and  $1.3\times$  on Fact (section 5.1). Moreover, lay-query performance transfers to the other query-type benchmarks while the reverse does not (section 5.2).

This work introduces **Deka Retrieval**, an LCR benchmark on the Thai Supreme Court Decisions (Deka)<sup>3</sup> together with a relevance-mining pipeline for curating LCR data from a civil-law jurisdiction. Our framework builds LCR pairs from judgments published online<sup>4</sup> through an LLM-assisted curation pipeline that decomposes each judgment into structured case components. For relevance mining, we curate a heuristic criterion over those components (Table 2) and drive the mining with an automated search agent whose candidates are verified by LLM judges. Beyond hard/easy positives and negatives, we further isolate a naturally occurring conflict judgment, the *anomaly* variant, which shares the anchor’s charge and facts yet reaches a different reasoning and decision (Table 2).

Our contributions are as follows:

1. Thai Supreme Court benchmark for LCR, produce another civil-law dataset based on case component criteria.
2. Study of asymmetric query gap in training and transfer performance across query types.
3. Empirical study of addition the conflict-judgment cases in dense-retrieval training.

## 2 Related Work

### 2.1 Legal Case Retrieval Datasets

LCR benchmark construction differs partly by legal tradition. In common-law systems, courts routinely

<sup>3</sup>We use *Deka* to refer to decisions of the Thai Supreme Court.

<sup>4</sup><https://deka.supremecourt.or.th>

cite prior cases with binding or persuasive authority, allowing citation links to serve as convenient, though incomplete, relevance signals. FIRE IRLeD and AILA tracks in Indian law (Mandal et al., 2017; Bhattacharya et al., 2019), COLIEE on Canadian case law (Kano et al., 2018; Goebel et al., 2023, 2025), and CLERC on U.S. case law (Hou et al., 2025) all build queries by masking a case’s citation sentence and treating the cited case as relevant. In civil-law systems, prior decisions generally lack the same formal precedential force, and case citations may be too sparse or incomplete to define relevance reliably. Chinese datasets LeCaRD (Ma et al., 2021), LeCaRDv2 (Li et al., 2024a), and CAIL2019-SCM (Xiao et al., 2019) grade case relevance through expert-designed criteria rather than citations (Feng et al., 2024). Most of these adopt a *symmetric* setting that uses a long-case *Fact* as a query (Li et al., 2023; Ma et al., 2021; Gao et al., 2024). Gao et al. (2024) move toward an asymmetric setting by compressing the *Fact* into a short query, and recent studies examine how layperson phrase legal search queries differently from legal experts (Zhao et al., 2026; Choi et al., 2026b; Paul et al., 2026).

## 2.2 Legal Case Retrieval Approaches

Neural retrievers for law adapt the dual-encoder paradigm (Deng et al., 2024b) to the length and terminology of judgments. Lawformer (Xiao et al., 2021) extends attention to long cases, while SAILER (Li et al., 2023) and DELTA (Li et al., 2025) inject document structure. The former reconstructs Reasoning and Decision from the *Fact*, the latter aligning key facts for a more discriminative encoder, both motivated by the observation that semantic similarity does not equal legal relevance. A second line injects explicit legal knowledge: CaseEncoder (Ma et al., 2023) enhances pre-training with legal signals, KELLER (Deng et al., 2024b) reformulates cases via charges and law articles, and other work incorporates legal elements directly (Deng et al., 2024a; Kim et al., 2025). Graph-based methods such as CaseLink (Tang et al., 2024) and LEXA (Tang et al., 2026) model case-to-case structure, and LLM-centric approaches use reasoning or agents for retrieval and reformulation (Tang et al., 2025; Tao et al., 2026).

## 2.3 LCR Cases Pair Mining

Relevance labels in existing LCR datasets rely on expert annotators scoring each axis numeri-

cally. LeCaRD used material and legal facts (Ma et al., 2021), then LeCaRDv2 used characterization, penalty, and procedure (Li et al., 2024a), a scheme Ma et al. (2024) later automate via LLM scoring of BM25-sampled pairs. Training-pair construction is typically driven by text similarity and shared charges: Deng et al. (2024b) label subfact pairs positive by shared charge and fall back to cosine similarity otherwise, while Gao et al. (2024) match charges and law articles exactly for positive augmentation. Such similarity-driven mining can be fragile as shown in Table 1 where cases can share a very similar event yet subtle factual differences can lead to different charges or decisions. This motivates Li et al. (2024b) to argue that reliable hard-negative selection requires domain knowledge of relevance rather than surface similarity.

## 3 Methodology

We scrape summarized judgments<sup>5</sup> of the Thai Supreme Court covering 2007–2025 and partition them chronologically<sup>6</sup>. Latest case from the past 6 years is considered as **Query set** and split by case type stratified sampling (Figure 7; 1,381 cases; 1,176 train, 100 evaluate, 105 test), each acting as an *anchor* that mines relevant cases from the older **Corpus set** (9,998 cases). This temporal split mirrors how a practitioner studies prior judgments to manage the case in hand and prevents future cases from leaking into the candidate pool. We further group case type manually for further case distribution analysis (Appendix A and Figure 6)

### 3.1 Curate Case Component

Since a judgment decompose into segment e.g. fact, judge reasoning, decision (Li et al., 2023; Ma et al., 2021; Li et al., 2024a), we decompose the summarized case passage into elements as shown in the first process of Figure 1. We use Gemini-3-flash-preview (Google, 2025) for extracting fields (*Fact*, *Reasoning*, and the list of *Charges*) the prompt is given in appendix A.1.1, and regular expressions for the *Section*, *Case Issue*, *Decision*. Unlike the symmetric setting that queries with the lengthy *Fact*, we curate two asymmetric queries per case, for a *layperson* (length under 100 character) and *legal-worker* role from the case fact (Gao et al., 2024), reflecting how each

<sup>5</sup>The official Supreme Court judgments are only available in summarized format.

<sup>6</sup>See case distribution per-year in Figure 2.

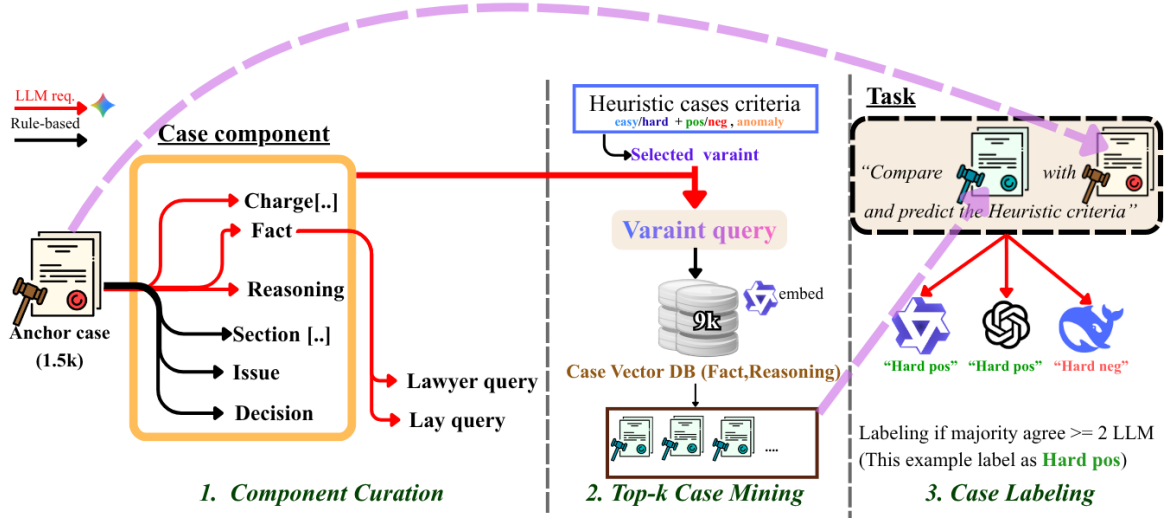


Figure 1: **Pipeline for Deka retrieval component curation and relevant case labeling.** (1) **Component Curation:** Generates each component from publicly summarized documents. (2) **Top- $k$  Case Mining:** Creates variant queries from the anchor context and applies mining criteria to retrieve relevant cases. (3) **Case Labeling:** Utilizes an LLM as a judge to evaluate the relevance level of mined cases relative to the anchor and assign labels.

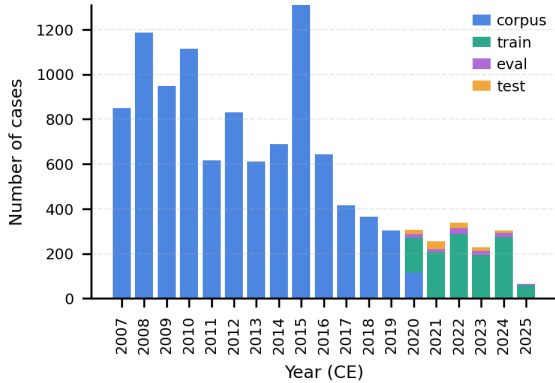


Figure 2: Temporal distribution of the Deka retrieval cases

| Criterion     | Charge | Fact         | Reasoning | Decision |
|---------------|--------|--------------|-----------|----------|
| Easy positive | Same   | Same         | Same      | Same     |
| Hard positive | Same   | Differ       | Same      | Same     |
| Hard negative | Differ | Same         | any       | any      |
| Easy negative |        | Non-relevant |           |          |
| Anomaly       | Same   | Same         | Differ    | Differ   |

Table 2: Heuristic case criteria across cases component.

role actually searches. All component texts are embedded with Qwen3-embedding-8B (Zhang et al., 2025) and indexed in a vector database for the mining stage.

### 3.2 Heuristic Case Criteria

Our heuristic for case relevance criteria jointly consider over the *Charge*, *Fact*, *Reasoning*, and *Decision* components in Table 2. An **easy positive**

agrees with the anchor on every component and is trivially relevant. A **hard positive** keeps the same charge, reasoning, and decision but differs in minor factual detail, so it is genuinely relevant yet looks dissimilar on the surface and is easily missed. A **hard negative** is an opposite: fact, reasoning, and decision resemble the anchor, but the charge differs, making it deceptively similar while not relevant. An **easy negative** is unrelated to the anchor overall. We further isolate a naturally occurring conflict judgment, which we call the *anomaly* variant: a case that shares the anchor’s charge and facts but reaches a different reasoning and decision.

### 3.3 Mining and Judging Cases

**Variant query generation.** To label candidates by their relevance to each anchor (second process in Figure 1), we generate one *variant query* per anchor and per criterion, using the same LLM as in Section 3.1 and setting prompt in appendix A.1.2. The model is prompted with the ‘*anchor case components*’ and the definition of the ‘*target criterion*’, so that the query is steered toward candidates standing in that particular relation to the anchor (Li et al., 2024b; Thakur et al., 2025; Choi et al., 2026a).

For example, if the *anchor case* is definition as figure 3, the *hard positive* variant difference only the anchor’s fact while charge, reasoning, and decision are similar, so that can represent in figure 4, where the underline marks the injected factual dif-

“Abduction and rape of a girl under 15;  
the court refused to suspend the prison  
sentence.”

Figure 3: Anchor’s context for variant query generation

ference.

“The defendant **arranged to meet the girl under 15**, then abducted and raped her;  
the court refused to suspend the prison  
sentence.”

Figure 4: Example of hard positive variant query generation from anchor (figure 3)

**Mining.** Each variant query searches the corpus over the *Fact* and *Reasoning* embedding spaces separately, returning a ranked candidate list. We then apply a charge filter on the *positives* and *anomaly* criterion to keep only candidates that share at least one charge with the anchor, applying a cosine-similarity check ( $\geq 0.96$ ) when an exact match isn’t possible, to prevent charge mismatches caused by grammatical variation in the LLM-generated text.

**Mined case selection.** We traverse the filtered candidates in ranked order and ask LLM judges to verify them (third process in Figure 1). Each judge sees only the *anchor case* and list of top-10 *candidate case*, without the criterion used to generate the query, and is asked to “compare the anchor and each candidate cases to predict the heuristic criterion”(prompt in appendix A.1.2). A candidate is accepted when the judges reach a majority and that majority label matches the criterion the query was generated for; otherwise it is discarded and we advance to the next candidate, stopping once one candidate is accepted for that criterion or the top-10 list is exhausted. Each criterion therefore contributes at most one candidate per anchor.

For the test set we use three judges, deepseek-chat-v3.1 (Liu et al., 2024), Qwen3-235b-a22b-2507 (Yang et al., 2025)<sup>7</sup>, and Gpt-5-mini (OpenAI, 2025), to reduce model-family bias. For the training set we use only Qwen3-235b-a22b-2507 as a judge for scalability.

<sup>7</sup><https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507>

## 4 Experimental Setup

### 4.1 Approach

Retrieve 9,998 case facts from **Corpus set** under three query type, the two asymmetric persona queries (**Lay**, **Lawyer**) and the **Fact** string as a symmetric reference (Ma et al., 2021).

**Baseline** We set BM25 as the baseline and utilize PyThaiNLP newmm tokenizer (Phatthiyaphaibun et al., 2023) for Thai language.

**Dense retrieval** We evaluate dense encoders in two settings. *Zero-shot*: the pretrained Qwen3-Embedding-0.6B (Zhang et al., 2025). *Fine-tuned*: starting from the same checkpoint, we train four objectives on our mined data. The mining stage gives each anchor four judged documents, one per criterion, and the objectives was selected from difference of sample require for training (Table 5). Configuration detail in Appendix B.

1. INFONCE (Oord et al., 2018; Karpukhin et al., 2020) is the baseline and uses polarity alone: one positive, with the anchor’s mined easy and hard negatives and all other in-batch documents treated as negatives of equal status. It consumes all four slots but distinguishes none of them, and fixes the level any graded objective has to beat.

2. WEIGHTED HARD-NEGATIVE adapts the  $\alpha$ -weighted term of supervised SimCSE (Gao et al., 2021), scaling the anchor’s own mined hard negative and remove easy negative

3. MULTI-POSITIVE modifies supervised contrastive learning (Khosla et al., 2020) by treating the hard positive as relevant alongside the easy one rather than merely being spared from the negative pool.

4. GRADED TRIPLET extends margin ranking loss (Schroff et al., 2015) to all four levels, giving easy negatives a larger margin than hard ones so the levels are encoded as distances rather than collapsed to polarity. It is the only objective with no in-batch pool every trained pair has been seen under our criteria

Legal PLMs pretrained on Chinese or English corpora (SAILER, DELTA, CaseEncoder) are excluded as trainable baselines, since our data is too small to re-pretrain.

### 4.2 Implementation Detail

We fully fine-tune across three NVIDIA V100 32 GB GPUs, and max token was setting to avoid text truncation (Figure 15). Asymmetric queries

(Lay, Lawyer) setting at 256 max tokens due to shorter queries, while the symmetric query and the corpus case facts use 1,300 max tokens. The *anomaly* variant asks what happens when relevance is defined by text and charge agreement alone, ignoring how the case was decided. Under the *anomaly* setup, we count the mined anomalies as positives as 89% of training anchors (1,058 of 1,176, Figure 11) contain anomalies, so counting them as positives is the cheapest way to enlarge the training set. We do this by duplicating the training batch and swapping the easy positive for the anomaly (Table 6). This setup was added to test whether including the anomalies as a training pair would result in retrieval degradation or not as this anomaly sometimes appeared in the real data distribution.

Early stop training on R@10 over the 100 evaluation anchors. Easy and hard positives count as relevant in evaluation and testing. The anomaly trained models additionally count anomalies as relevant during evaluation but the test-set never does for compare with normally training.

## 5 Results

### 5.1 Performance Under Symmetric Setup

Table 3 shows retriever performance. Under a symmetric setting, where the same input type is used for training and evaluation, retrievers trained and evaluated on *Fact* contexts generally achieve the strongest performance. However, simpler query type such as *Lawyer* and *Lay* shows lower scores across all metrics. This result is consistent with the intuition that more information-rich inputs are easier to align with the relevance labels during training. BM25 remains a competitive baseline: it outperforms zero-shot Qwen3-Embedding-0.6B on the *Lawyer* queries, although the dense retriever performs better on the *Lay* queries and *Fact* contexts. Across the standard fine-tuning objectives, in-domain fine-tuning generally yields substantial improvements over the zero-shot dense baseline. However, no single objective consistently performs best across all input types and evaluation metrics. Among the models trained without anomaly augmentation, the INFONCE-trained retriever achieves the best performance on four of the nine evaluation settings: R10 and NDCG10 for *Lay* queries, and NDCG10 and MRR for *Fact* contexts. Nevertheless, its performance remains broadly competitive with that of WEIGHTED and MULTIPOS.

### 5.2 An asymmetric setup Shows Opposite Direction

Table 3 shows the results under the asymmetric setup, in which the input type used for training differs from that used for evaluation. The results indicate that matching the training input type to the test input type does not always produce the best retriever. Across objectives, retrievers trained on *Lay queries* often remain competitive when evaluated on the more information-rich *Lawyer queries* and *Fact contexts*. Retriever trained on *Lawyer queries* also often transfer well to *Fact contexts*, whereas their performance on *Lay queries* is consistently lower than that obtained by training directly on *Lay queries*. In contrast, retrievers trained on *Fact contexts* generally transfer poorly to *Lay queries*. These results reveal an asymmetric pattern: **training on less informative queries can generalize well to more informative inputs, whereas transfer in the opposite direction is less effective.** We hypothesize that this behavior arises because training on *Lay queries* presents a harder alignment problem, as these queries retain only the minimal information needed to express the information need. This constraint may force the retriever to learn more robust representations of relevance. By contrast, *Fact contexts* contain substantially more information, allowing the retriever to rely on cues that may be unavailable in shorter queries. This reliance may result in poorer generalization to less informative *Lawyer* and *Lay queries*.

### 5.3 Anomaly cases addition effect

As courts sometimes have anomalous cases where the same fact and charge yet final decision differ (see Table 2). We conduct an experiment to see what happens if we also incorporate these anomalies into the training data, reflecting the noise in the real-world distribution. This way we almost double the positives of all cases while, enrich the training data with noise.

According to the result in Table 3, the trade is usually not worth taking. Of 36 evaluation cells (4 objectives  $\times$  3 training query types  $\times$  3 test query types), 27 fall, 7 rise, and 2 are unchanged relative to the normally fine-tuned counterpart. The added pairs teach the retriever to rank conflicting judgments alongside true positives but the test set counts only easy and hard positives as relevant. The extra data acts as label noise and the larger training set does not compensate for it. The retriever cannot

| Objective      | Train on | An. | Lay query    |              |              | Lawyer query |              |              | Fact context |              |              |
|----------------|----------|-----|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                |          |     | R@10         | NDCG@10      | MRR          | R@10         | NDCG@10      | MRR          | R@10         | NDCG@10      | MRR          |
| BM-25          | -        |     | 6.19         | 4.31         | 6.6          | 22.86        | 15.23        | 18.43        | 26.19        | 16.34        | 20.48        |
| ZERO-SHOT      | -        |     | 7.62         | 4.58         | 7.45         | 20.0         | 11.94        | 15.46        | 26.67        | 17.46        | 24.63        |
| INFONCE        | Lay      |     | <b>19.52</b> | <b>11.83</b> | <b>13.79</b> | 26.19        | <b>16.05</b> | <b>18.64</b> | 31.90        | 21.09        | <b>28.27</b> |
|                | Lawyer   |     | 12.86        | 8.89         | 12.31        | <b>27.14</b> | 15.73        | 17.92        | <b>34.76</b> | <b>23.06</b> | 27.06        |
|                | Fact     |     | 8.57         | 5.23         | 7.39         | 25.24        | 14.53        | 16.33        | 33.81        | 22.43        | 27.57        |
|                | Lay      | ✓   | 15.71        | 9.63         | <b>12.25</b> | 24.29        | 15.87        | <b>19.64</b> | 28.10        | 18.65        | <b>23.88</b> |
|                | Lawyer   | ✓   | <b>16.19</b> | <b>10.06</b> | 11.02        | 24.76        | <b>16.33</b> | 19.20        | <b>24.76</b> | 17.64        | 22.70        |
|                | Fact     | ✓   | 10.48        | 6.35         | 8.56         | 21.43        | 13.55        | 15.67        | <b>29.05</b> | <b>20.02</b> | 23.65        |
| WEIGHTED       | Lay      |     | <b>16.19</b> | <b>10.72</b> | <b>14.57</b> | <b>26.67</b> | <b>16.01</b> | <b>18.55</b> | 31.90        | 19.81        | 26.62        |
|                | Lawyer   |     | 13.33        | 8.90         | 11.81        | <b>26.67</b> | 15.37        | 18.06        | <b>32.86</b> | 21.94        | 26.25        |
|                | Fact     |     | 8.57         | 5.41         | 7.65         | 24.76        | 14.26        | 15.97        | <b>32.86</b> | <b>21.95</b> | <b>27.14</b> |
|                | Lay      | ✓   | <b>15.24</b> | <b>8.21</b>  | <b>9.98</b>  | <b>23.33</b> | <b>14.04</b> | <b>15.36</b> | 28.10        | 18.15        | 22.32        |
|                | Lawyer   | ✓   | 11.90        | 7.14         | 8.27         | 21.90        | 13.61        | 15.03        | <b>28.57</b> | <b>19.53</b> | <b>23.73</b> |
|                | Fact     | ✓   | 8.57         | 4.95         | 6.90         | 20.48        | 12.53        | 14.96        | 28.10        | 19.29        | 23.52        |
| MULTIPOS       | Lay      |     | <b>16.67</b> | <b>9.60</b>  | <b>13.23</b> | 25.24        | 16.34        | <b>20.31</b> | 30.95        | 20.30        | 25.65        |
|                | Lawyer   |     | 13.33        | 7.93         | 9.53         | <b>29.52</b> | <b>17.18</b> | 19.74        | <b>32.38</b> | <b>22.88</b> | <b>29.09</b> |
|                | Fact     |     | 11.90        | 6.26         | 6.52         | 21.90        | 13.57        | 16.64        | 31.43        | 20.88        | 23.87        |
|                | Lay      | ✓   | <b>21.43</b> | <b>11.47</b> | <b>12.66</b> | 21.43        | 15.09        | <b>19.28</b> | 28.57        | 18.56        | 23.09        |
|                | Lawyer   | ✓   | 12.38        | 7.03         | 8.83         | 25.24        | 14.80        | 16.87        | 27.62        | 18.06        | 23.68        |
|                | Fact     | ✓   | 9.52         | 5.77         | 8.05         | <b>26.19</b> | <b>15.51</b> | 18.09        | <b>33.33</b> | <b>22.11</b> | <b>26.60</b> |
| GRADED TRIPLET | Lay      |     | <b>12.86</b> | <b>7.73</b>  | <b>10.74</b> | <b>25.24</b> | <b>15.60</b> | 18.46        | 29.52        | 19.71        | 25.46        |
|                | Lawyer   |     | 11.90        | 6.21         | 8.24         | 24.76        | 14.34        | 16.68        | 32.86        | <b>22.16</b> | <b>27.40</b> |
|                | Fact     |     | 10.95        | 5.88         | 7.41         | 23.33        | 14.99        | <b>18.64</b> | <b>34.76</b> | 21.88        | 27.01        |
|                | Lay      | ✓   | <b>12.38</b> | <b>8.16</b>  | <b>11.72</b> | <b>26.67</b> | <b>14.15</b> | 15.46        | 30.00        | <b>19.22</b> | <b>24.19</b> |
|                | Lawyer   | ✓   | 11.90        | 6.53         | 8.20         | 20.00        | 11.19        | 13.93        | 28.10        | 17.96        | 21.18        |
|                | Fact     | ✓   | 10.00        | 5.82         | 7.15         | 21.43        | 13.32        | <b>16.45</b> | <b>30.48</b> | 18.89        | 22.56        |

Table 3: Main performance on Deka Retrieval: zero-shot and fine-tuned retrievers evaluated on each query type. **Lay** is a short everyday query with no legal terms, **Lawyer** is a query written in legal terms, and **Fact** is the full case-fact passage. Cyan cell indicate settings trained and tested on the same query type

learn the distinction from what it is given: a positive and an anomaly share the anchor’s facts and charge, and differ only in reasoning and decision, which the candidate’s Fact does not carry. Figure 13 helps explain why explicit anomaly supervision is needed. Before task-specific fine-tuning, anomaly cases have similarity and overlap values close to those of positive cases across several legal components. The pretrained encoder therefore does not clearly distinguish anomaly cases from positive cases on its own.

## 6 Ablation Study

### 6.1 Reranking as Post-processing

A reranker is commonly applied as a second stage to improve the ranking produced by a first-stage retriever (Nogueira and Cho, 2019; Rosa et al., 2022), including in legal case retrieval (Ngo et al., 2026). To determine whether reranking also improves fine-tuned dense retrievers, we rerank the top-100 retrieved cases using four rerankers: Qwen3-Reranker-0.6B, Qwen3-Reranker-8B, bge-reranker-v2-m3<sup>8</sup>, which is based on BGE-M3 (Chen et al., 2024), and the Gemma-2B-based bge-reranker-v2-gemma<sup>9</sup>. Figures 16, 17, and 18 show the changes produced by these rerankers

<sup>8</sup><https://huggingface.co/BAAI/bge-reranker-v2-m3>

<sup>9</sup><https://huggingface.co/BAAI/bge-reranker-v2-gemma>

in R10, MRR, and NDCG10, respectively.

Across the three retrieval metrics, Qwen3-Reranker-8B produces the strongest and most consistent improvements. It improves R@10 and NDCG@10 in every setting and improves MRR in all but two settings, where the decreases are only 0.1 percentage points. The Gemma-based BGE reranker generally improves R10, but its effects on MRR and NDCG@10 are more dependent on the first-stage retriever and test input type. The smaller bge-reranker-v2-m3 and Qwen3-Reranker-0.6B models show more mixed results, particularly for *Fact* contexts. These findings indicate that **Qwen3-Reranker-8B is the most robust reranker across first-stage retrievers and input types in our legal case retrieval setting.**

Reranking also partially offsets the performance loss caused by anomaly training. Under Qwen3-Reranker-8B, the anomaly-trained model obtains a larger average R10 improvement than its non-anomaly counterpart in 9 of the 12 objective-training-input pairs, where each improvement is averaged over the three test input types. This result is consistent with Section 5.3, which shows that anomaly-trained retrievers place more anomaly-labeled cases in the top 100. The cross-encoder can reassess these candidates and demote those that are not relevant.

After reranking with Qwen3-Reranker-8B, the

performance differences among BM25, the zero-shot dense retriever, and the fine-tuned retrievers become smaller for *Lawyer* queries and *Fact* contexts. For example, the reranked zero-shot retriever reaches 36.19 R10 on *Lawyer* queries, making it competitive with many of the reranked fine-tuned models. The advantage of first-stage fine-tuning remains clearest for *Lay* queries. This pattern is consistent with the asymmetry described in Section 5.1: task-specific fine-tuning contributes most when the query is expressed in short, everyday language, whereas a strong reranker can compensate for much of the first-stage difference when the input already contains legal terminology or detailed facts.

## 6.2 Relevance case mining report

At first, we curate a 296 sample with a stratified partition that preserves the proportion of each case type in the test set. After running the mining in Section 3.3, however 191 of the 296 test cases produced an incomplete mining result. Figure 9 plots the case-type distribution of these incomplete anchor. As the bar chart shows, they are dominated by financial cases such as Property, Contract & Debt, and Loan & Guarantee. We attribute that financial cases narratives are brief and structurally near-identical (e.g., a party borrows a sum at an agreed rate and then defaults), so two cases differ only in fine legal elements such as the guarantee type or the interest-rate band rather than in the fact narrative itself, this challenge is frame in private-lending task (Xiao et al., 2019).

## 6.3 Limited of candidate cases context

Following the standard setting in prior LCR work (Ma et al., 2021; Xiao et al., 2019; Bhattacharya et al., 2019), our retriever and reranker observe only the *Fact* section of each candidate case. As Table 2 shows, anomaly cases are difference from positives by their *Reasoning* and *Decision* which are never seen by retriever. This gap explaining why anomaly augmentation degrades performance. Including reasoning/decision context in retrieve and training to learn case-component level is left as future work.

## 7 Conclusion

We introduced Deka Retrieval, an LCR benchmark on Thai Supreme Court decisions, and a mining pipeline that stakes relevance over case components rather than over surface similarity. The criteria

make two properties of LCR measurable. The first query type is retrieving different results; lengthy text contains legal terms, while laypeople’s queries do not. This gap is also asymmetric, since training on the lay rise score the most and perform well on the longer query benchmark, while the reverse fails. Second, conflicting judgments are not free supervision; counting them as positives lowers performance for 27 of 36 trained models. We hypothesize that a civil-law corpus such as Thai supreme court, without citation links may contain such conflicting pairs by construction, and that a pipeline defining relevance through text and charge agreement alone would have no signal by which to separate them.

## Acknowledgement

This work was conducted while the first author was an intern at VISAI. We thank Vidyasirimedhi Institute of Science and Technology (VISTEC) for providing the computing resources used in this research.

## Limitations

**Labels are LLM-derived and unaudited by legal experts** Case components, variant queries, and relevance labels all come from LLMs, the test set uses three judges from different model families. Human expert agreement is not explored, so the criteria are validated by construction.

**Source text is summarized** The publicly available Deka contains only long/short summaries, which retain the legal issues, facts, and rulings but omit witness lists and detailed circumstances. Retrieval the summarized text can overlook case detail that not publicly in website.

**Model and compute scope** Data curate and evaluate training objective was adopt retriever by only Qwen3-Embedding which can be causing a bias retrieve. Utilize cross model family is left to future work.

**Test-set cannot evaluate every cases type** Section 6.2 report that there is a financial and property case type unsuccessful mining. Therefore we exclude this group in test-set and report the challenge of define heuristic criteria on financial domain.

## References

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhat-

- tacharya, and Prasenjit Majumder. 2019. Overview of the fire 2019 aila track: Artificial intelligence for legal assistance. In *FIRE (working notes)*, pages 1–12.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Hyewon Choi, Jooyoung Choi, Hansol Jang, Hyun Kim, Chulmin Yun, ChangWook Jun, and Stanley Jungkyu Choi. 2026a. Arhn: Answer-centric relabeling of hard negatives with open-source llms for dense retrieval. *arXiv preprint arXiv:2604.11092*.
- Janghyeok Choi, Jaewon Lee, and Sungzoon Cho. 2026b. Daldall: Data augmentation for lexical and semantic diverse in legal domain by leveraging llm-persona. *arXiv preprint arXiv:2603.22765*.
- Somaiyeh Dehghan and Mehmet Fatih Amasyali. 2022. Supmpn: supervised multiple positives and negatives contrastive learning model for semantic textual similarity. *Applied Sciences*, 12(19):9659.
- Chenlong Deng, Zhicheng Dou, Yujia Zhou, Peitian Zhang, and Kelong Mao. 2024a. An element is worth a thousand words: Enhancing legal case retrieval by incorporating legal elements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2354–2365.
- Chenlong Deng, Kelong Mao, and Zhicheng Dou. 2024b. Learning interpretable legal case retrieval via knowledge-guided case reformulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1253–1265.
- Yi Feng, Chuanyi Li, and Vincent Ng. 2024. [Legal case retrieval: A survey of the state of the art](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6472–6485, Bangkok, Thailand. Association for Computational Linguistics.
- Cheng Gao, Chaojun Xiao, Zhenghao Liu, Huimin Chen, Zhiyuan Liu, and Maosong Sun. 2024. Enhancing legal case retrieval via scaling high-quality synthetic query-candidate pairs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Calum Kwan, Ken Satoh, Hiroaki Yamada, and Masaharu Yoshioka. 2025. An overview of the coliee 2025 competition: Legal case law and statute law information retrieval and entailment. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*, pages 506–515.
- Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. Summary of the competition on legal information, extraction/entailment (coliee) 2023. In *Proceedings of the nineteenth international conference on artificial intelligence and law*, pages 472–480.
- Google. 2025. Gemini 3 flash preview. <https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview>. Accessed: 2026-08-19.
- Abe Bohan Hou, Orion Weller, Guanghui Qin, Eugene Yang, Dawn Lawrie, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. 2025. Clerc: A dataset for us legal case retrieval and retrieval-augmented analysis generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7898–7913.
- Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. 2018. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In *JSAIL International symposium on artificial intelligence*, pages 177–192. Springer.
- 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.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.
- Chaeun Kim, Jinu Lee, and Wonseok Hwang. 2025. Legalsearchlm: Rethinking legal case retrieval as legal elements generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4521–4554.
- Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. Sailer: structure-aware pre-trained language model for legal case retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1035–1044.

- Haitao Li, Qingyao Ai, Xinyan Han, Jia Chen, Qian Dong, and Yiqun Liu. 2025. Delta: Pre-train a discriminative encoder for legal case retrieval via structural word alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27072–27080.
- Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024a. Lecardv2: A large-scale chinese legal case retrieval dataset. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2251–2260.
- Xiaopeng Li, Xiangyang Li, Hao Zhang, Zhaocheng Du, Pengyue Jia, Yichao Wang, Xiangyu Zhao, Huifeng Guo, and Ruiming Tang. 2024b. Syneg: Llm-driven synthetic hard-negatives for dense retrieval. *arXiv preprint arXiv:2412.17250*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Shengjie Ma, Qi Chu, Jiaxin Mao, Xuhui Jiang, Haozhe Duan, and Chong Chen. 2024. Leveraging large language models for relevance judgments in legal case retrieval. *arXiv preprint arXiv:2403.18405*.
- Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021. Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.
- Yixiao Ma, Yueyue Wu, Weihang Su, Qingyao Ai, and Yiqun Liu. 2023. Caseencoder: A knowledge-enhanced pre-trained model for legal case encoding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7134–7143.
- Arpan Mandal, Kripabandhu Ghosh, Arnab Bhat-tacharya, Arindam Pal, and Saptarshi Ghosh. 2017. Overview of the fire 2017 irted track: Information retrieval from legal documents. In *FIRE (Working Notes)*, pages 63–68.
- Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. 2024. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*.
- Thuong-Hieu Ngo, Hoang-Trung Nguyen, Huu-Dong Nguyen, Xuan-Bach Le, Le-Dung Nguyen, Quang-Thanh Tran, Ha-Thanh Nguyen, and Thi-Hai-Yen Vuong. 2026. Nowj@ coliee 2026: Adaptive pipelines for legal retrieval and reasoning. *arXiv preprint arXiv:2607.16603*.
- Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- OpenAI. 2025. Gpt-5 mini (version gpt-5-mini-2025-08-07). <https://chatgpt.com>. Large language model. Accessed: 2026-07-17.
- Shounak Paul, Raghav Dogra, Pawan Goyal, and Saptarshi Ghosh. 2026. Ilsic: Corpora for identifying indian legal statutes from queries by laypeople. *arXiv preprint arXiv:2602.00881*.
- Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *European conference on information retrieval*, pages 397–412. Springer.
- Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntornitip, 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, Singapore. Association for Computational Linguistics.
- Guilherme Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. In defense of cross-encoders for zero-shot retrieval, december 2022. URL <https://arxiv.org/abs/2212.06121>.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. *Simple entity-centric questions challenge dense retrievers*. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yunqiu Shao, Haitao Li, Yueyue Wu, Yiqun Liu, Qingyao Ai, Jiaxin Mao, Yixiao Ma, and Shaoping Ma. 2023. An intent taxonomy of legal case retrieval. *ACM Transactions on Information Systems*, 42(2):1–27.
- Lynda Tamine and Lorraine Goeuriot. 2021. Semantic information retrieval on medical texts: Research challenges, survey, and open issues. *ACM Computing Surveys (CSUR)*, 54(7):1–38.
- Yanran Tang, Ruihong Qiu, Xue Li, and Zi Huang. 2025. Reakase-8b: Legal case retrieval via knowledge and reasoning representations with llms. In *Australasian Database Conference*, pages 306–320. Springer.

Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2026. Lexa: Legal case retrieval via graph contrastive learning with contextualised llm embeddings. *World Wide Web*, 29(2):20.

Yanran Tang, Ruihong Qiu, Hongzhi Yin, Xue Li, and Zi Huang. 2024. Caselink: Inductive graph learning for legal case retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2199–2209.

Mingxu Tao, Jiawei Hu, Xian Zhou, Wenpeng Hu, Jiajun Cheng, Yunbo Cao, Zhunchen Luo, and Guotong Geng. 2026. When rules learn: A self-evolving agent for legal case retrieval. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 43348–43365.

Nandan Thakur, Crystina Zhang, Xueguang Ma, and Jimmy Lin. 2025. Hard negatives, hard lessons: Re-visiting training data quality for robust information retrieval with llms. *arXiv preprint arXiv:2505.16967*.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Tianyang Zhang, Xianpei Han, Zhen Hu, Heng Wang, et al. 2019. Cail2019-scm: a dataset of similar case matching in legal domain. *arXiv preprint arXiv:1911.08962*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. *Qwen3 technical report*. *Preprint*, arXiv:2505.09388.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1503–1512.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Ethan Zhao, Maksym Taranukhin, Wei Cui, Moira Aikenhead, and Vered Shwartz. 2026. Canlegalrag-bench: Evaluating retrieval-augmented generation on canadian case law. *arXiv preprint arXiv:2605.30497*.

## A Data statistic



“ The defendant, the mother’s ex-partner, ... had raised the victim (a girl aged 11) like a father ... so that she was under his authority. ... while the mother had left her in his care, he ... inserted his finger into her genitals and attempted penetration ... occurring 4 times in total. Later ... he slapped and kicked her until injured ”



“ The defendant, holding the status of a caregiver, exploited the fact that the 11-year-old victim was under his care and repeatedly penetrated the victim with his finger and his own genitals, and also used physical force causing injury, while the victim was under his custodial authority ”



“ Someone that mother knows did something bad to a kid and beat her up too ”

Figure 5: Examples of query inputs in child sexual assault cases, categorized by case facts, lawyer queries, and layperson queries

### A.1 LLM Prompt

#### A.1.1 Curate case component

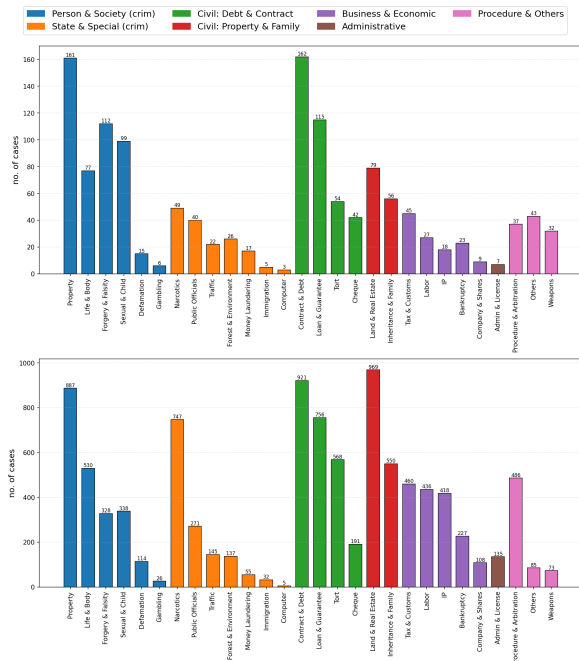


Figure 6: Clustering case type in query and corpus, query-set(top) and corpus-set(bottom)

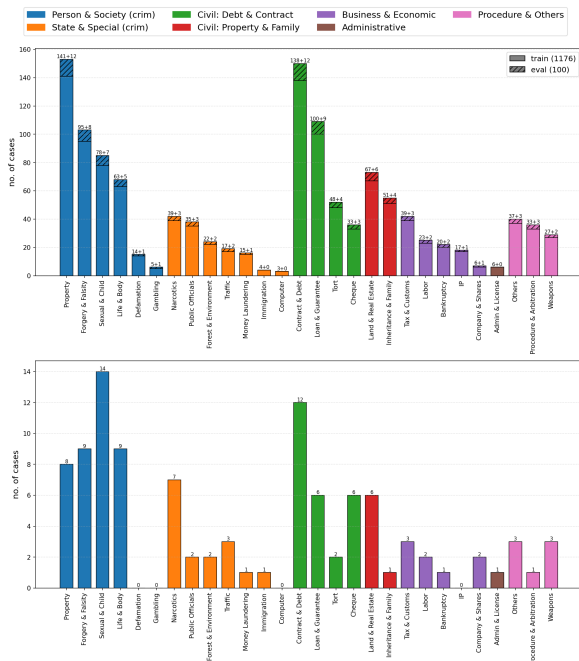


Figure 7: Clustering case type in Query-set: compare train and eval(top) and test set(bottom)

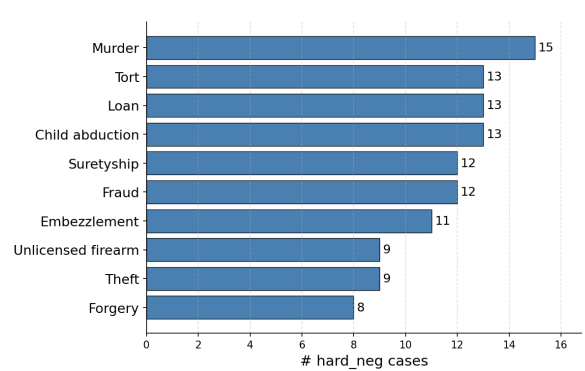


Figure 8: Top frequency charge overlap between anchor and hard negative cases in test-set.

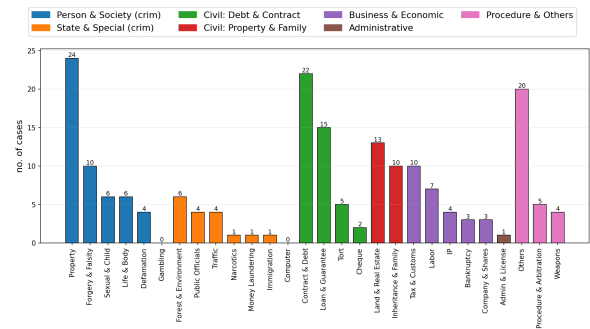


Figure 9: The distribution of test-set cases across classes became uneven, as some were dropped due to the difficulty of mining them

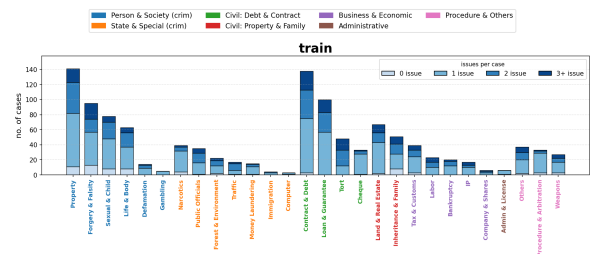


Figure 10: The distribution of case-issue from cases regex extraction across case-types

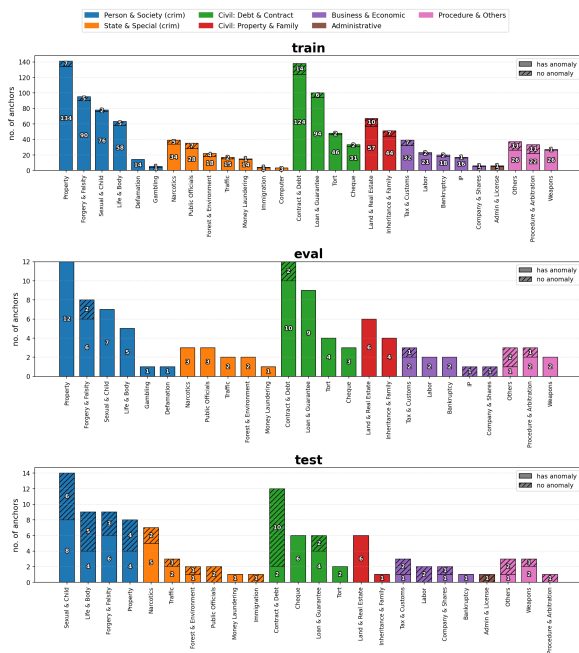


Figure 11: The distribution of mining anomaly cases varies across case types

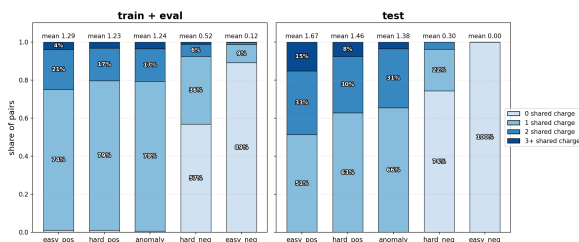


Figure 12: the distribution of charge overlap across criteria variants : test-set distribution visualizes charge relevance with anchor charge amounts across each criteria variant based on majority judge consensus

#### Anchor — Drink driving, Negligent Homicide

|           |                                                                                                                                                                                              |
|-----------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fact      | Driving while drunk, the defendant lost control and caused a crash that killed another road user on the spot.                                                                                |
| Reasoning | The charge is causing death by <b>negligence</b> , together with drink-driving; the intoxicated, careless driving was the direct cause of the fatal accident, with no intent to harm anyone. |
| Decision  | Judgment <b>affirmed</b> : <b>guilty</b> of negligent homicide, one year after a guilty plea.                                                                                                |

#### Hard Negative — Drink driving, Attempted Murder

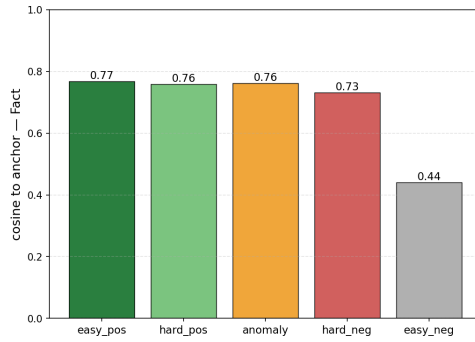
|           |                                                                                                                                                               |
|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fact      | Enraged after a horn exchange, the drunk defendant chased a motorcycle, deliberately rammed it, and dragged it beneath his car, gravely injuring both riders. |
| Reasoning | The charge is <b>murder</b> ; deliberately ramming with a far heavier vehicle showed intent to kill, and voluntary intoxication was no defence.               |
| Decision  | Judgment <b>modified</b> : <b>guilty</b> of attempted murder together with drink-driving, eleven years in total.                                              |

Table 4: Pair of drink-driving anchor and hard negative case

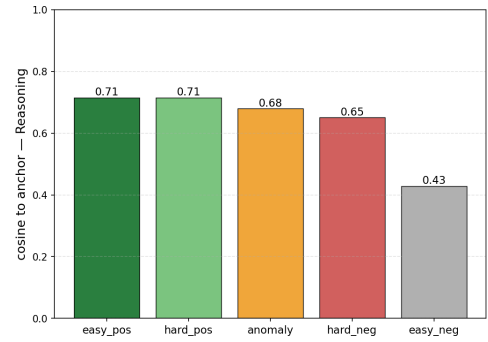
#### Fact generation

You are a legal expert. Summarize the "factual circumstances" of the entire case into a single passage. Rules: - Replace all personal names with roles, e.g. defendant, victim, deceased, co-offender - Remove the court's legal reasoning and verdict - Focus only on the circumstances that constitute the core elements of the offense: state the motive, process, outcome, time, and place of the defendant's conduct - Cover every charge in the case within the single passage, in chronological order Answer as a single passage (one paragraph) – no labels, headings, markdown, or bullet lists. Example: "The defendant deceived the victim by making false statements, contacting the victim directly via a messaging app, persuading the victim to co-invest in an imported fashion clothing business, while showing forged purchase-order document photos and guaranteeing a 15

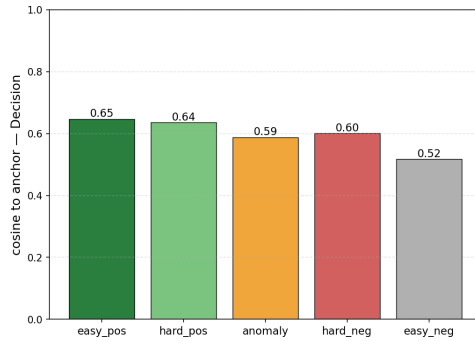
User message: Case text: {long\_text}



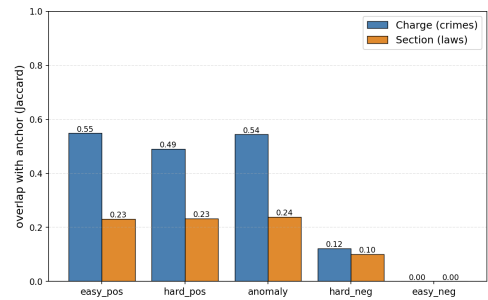
(a) *Fact* (cosine)



(b) *Court Reasoning* (cosine)



(c) *Final Decision* (overlap)



(d) *Charge/Section* (overlap  $\geq 1$ )

Figure 13: Distribution of pairwise similarity between anchor and candidate cases in **test-set**, across the four legal case components used in our relevance criteria. *Fact* (a) and *Court Reasoning* (b) are measured by Qwen3-Embedding-8B embedding cosine similarity, whereas *Final Decision* (c) and *Charge/Section* (d) are measured by token/label overlap.

### Charge generation

Read the Supreme Court judgment and extract the core "charges (CRIMES)" – every one, in Thai, separated by ; - Use short charge names with no modifiers, e.g. "theft" not "jointly committing theft by..." - No duplicates – if the same charge applies across multiple counts, write it only once  
Answer in a single line only: CRIMES: <charge1>;<charge2>

**User message:** Judgment: {long\_text}

### Reasoning generation

You are a Thai legal expert. Read the full Supreme Court judgment and summarize "the reasoning the court used to decide the case" (legal fact). Rules: - Use only the Supreme Court's ruling. If the case went through multiple court levels, consider only the most recent (highest) level's ruling. - Anonymize proper names such as names

of people, organizations, places. - Include only what the court actually ruled in the judgment – do not reason on your own. If the court's decision goes against common sense, write it exactly as the court decided. - Write it as applying the facts to the elements of the law: which conduct does/does not fall under which section, and why. - Always state the case's main charge/dispute, even if the Supreme Court did not argue that point (already settled in lower courts) – then follow with the point(s) the Supreme Court actually decided. Output only the legal fact text – no heading, no preamble, no numbering.

**User message:** {long\_text}

### Lay query generation

You are an "ordinary person" with no legal knowledge, typing a search query about a case you encountered. From the given case

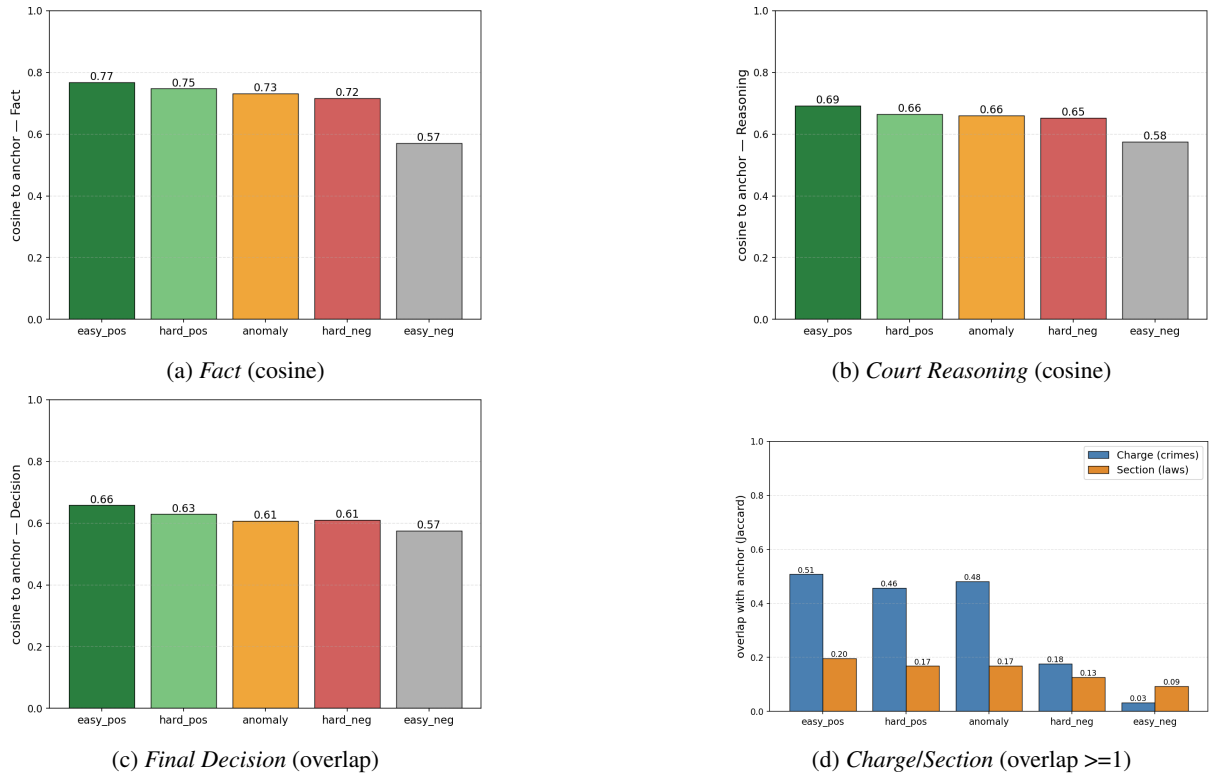


Figure 14: Distribution of pairwise similarity between anchor and candidate cases in **train** and **eval-set**, across the four legal case components used in our relevance criteria. *Fact* (a) and *Court Reasoning* (b) are measured by Qwen3-Embedding-8B embedding cosine similarity, whereas *Final Decision* (c) and *Charge/Section* (d) are measured by token/label overlap.

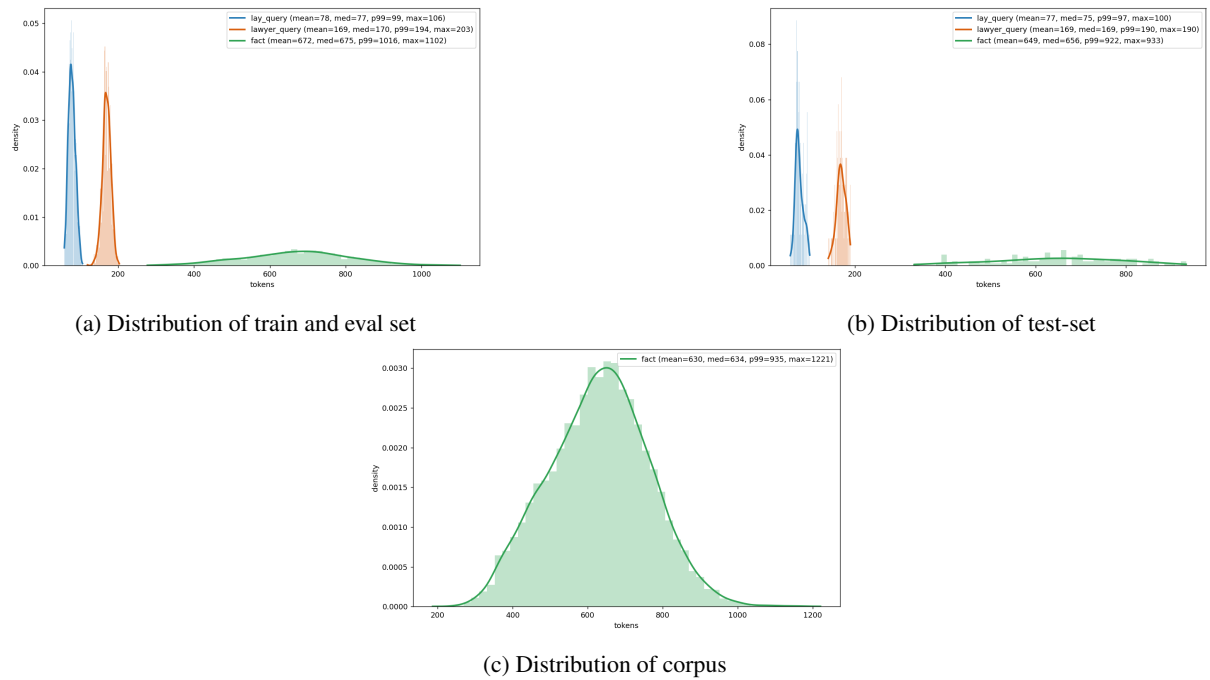


Figure 15: Distribution of token counts using the *qwen3-embedding-8b* tokenizer

facts, produce a “search query” the way a layperson would.

Principle: ordinary people only remember the “surface-level circumstances” and don’t know which details matter legally, so they tend to unknowingly drop them. “Collapse or omit” the following decisive elements: - Exact age → use a broad term (child/teenager/adult), don’t state whether it is over/under a legal threshold - The legal relationship between the parties (parent/guardian, employer) → omit or state broadly - Consent/non-consent → do not state it - Whether acting alone or with others → do not distinguish - Intent/negligence, use of force/weapons → do not state unless an ordinary person would mention it  
Format: - Very short, 1 declarative sentence, spoken/colloquial language - Remove proper names (people/places/brands/dates) → replace with generic terms - No legal jargon, no charge names or section numbers - No questions/requests, and do not end with polite particles (khrap/kha)

**User message:** Case facts: {fact}

#### Lawyer query generation

You are a “legal worker” (lawyer or law clerk) searching for prior cases relevant to a matter in hand. From the given case facts, produce a “search query” the way a legal professional would. Principle: legal workers search by the decisive legal elements, because they know which details determine the charge and the outcome. “Preserve and make explicit” the following elements: - Exact age relative to the legal threshold (e.g. under 15) when relevant - The legal relationship between the parties (parent/guardian, employer, custodian, lessor/lessee) - Consent / non-consent - Acting alone or jointly with others - Intent vs. negligence, and use of force or weapons  
Format: - Concise but complete, using precise legal terminology - Refer to conduct by its legal characterisation where clear (e.g. embezzlement, theft, negligent homicide) - Remove proper names (people/places/brands/dates) -> re-

place with generic terms - No questions/requests, and do not end with polite particles (khrap/kha)

**User message:** Case facts: {fact}

#### A.1.2 Mining and Judging Cases

##### Variant definition

easy\_pos: same facts, same charge, same court reasoning and same verdict  
hard\_pos: slightly different facts, same charge, but same court reasoning and same verdict  
anomaly: same facts, same charge, but different court reasoning/verdict  
hard\_neg: same facts/circumstances, but clearly decided under a different charge  
easy\_neg: completely unrelated, a different matter entirely

##### Variant query generation

You are a retrieval assistant for Supreme Court judgments. Design “semantic search queries” from the main case (Anchor).

{Variant definition}

Do the following:

- 1) Pick exactly one type from the remain list above (the group you think you can write the best-matching queries for)
- 2) Generate 5-8 search queries for that one chosen type only -- every query's target must be the same single group, never mixed

Each query must choose a field: fact (event/circumstances) | legal reasoning (legal principle)

Answer as a JSON array only: [{"field": "fact", "query": "...", "target": "<chosen group>"}]

**User message:**

{anchor\_context}

**{anchor\_context} field format:**

case no.: {deka}  
charge(s): {charge}  
law section(s): {laws}  
facts: {fact}  
issue(s): {issue}  
legal reasoning: {reasoning}  
verdict: {decision}

### Mined case selection

You are a Thai judge. Compare the "main case (Anchor)" against a "comparison case (Candidate)" and classify which group it belongs to.

Classify into 1 of 5 groups -- decide using 3 axes: (a) facts (b) charge (c) court reasoning/verdict

{Variant definition}

The axis that separates anomaly from hard\_pos = the facts -- facts that match almost entirely but the court ruled differently = anomaly (a self-contradicting body of precedent). If the facts have a difference that explains the different verdict = hard\_pos.

If the "charge is the same" it must always be easy\_pos / hard\_pos / anomaly -- never classify it as hard\_neg just because the case was decided on a procedural or secondary issue. hard\_neg must come from "a different charge" only.

Answer as a JSON array only, covering every id given, with no other fields or explanation:  
[{"id":1,"quadrant":"easy\_pos"}]

### User message:

```
=== Anchor ===
{anchor_context}
=== List of comparison case ===
[ {candidate1_context}
.
.
.
{candidate10_context} ]
```

### A.1.3 Inference LLM

#### Embedding instruction

Given a legal case search query, retrieve relevant prior Supreme Court cases

#### User message:

[Your text goes here]

#### Reranker instruction

Given a legal case search query, retrieve prior Supreme Court cases whose facts match the query

| Configuration  | Data batch                       |
|----------------|----------------------------------|
| InfoNCE        | $c_{pe}, c_{nh}, c_{ne}$         |
| Weighted       | $c_{pe}, c_{nh}$                 |
| Multi-positive | $c_{pe}, c_{ph}, c_{nh}, c_{ne}$ |
| Graded triplet | $c_{pe}, c_{ph}, c_{nh}, c_{ne}$ |

Table 5: The training batch  $pe/ph$  = easy/hard positive and  $ne/nh$  = easy/hard negative

### Query:

[Your text goes here]

### Fact:

[Your text goes here]

## B Training Objectives

### B.1 Shared setup

All configurations share one backbone, batching scheme, and optimiser, and differ only in which mined slots an anchor carries and how the loss consumes them. Queries and documents pass through the same encoder, with last-token pooling over left-padded inputs followed by  $L_2$  normalisation; a task instruction is prepended on the query side only (Appendix A.1.3).

Each anchor training with  $L_2$ -normalised embeddings,  $s_{ij} = \mathbf{q}_i^\top \mathbf{d}_j$  and  $S_{ij} = s_{ij}/\tau$  with  $\tau = 0.05$ , where  $i$  indexes anchors and  $j$  indexes slots in the pool. Because the pool is shared, a document mined as a positive for anchor  $i$  can appear as another anchor’s slot and be scored as a negative for  $i$ . Before every softmax we set  $\tilde{S}_{ij} = -\infty$  when  $\text{docid}(j) \in \mathcal{R}_i$  and  $j \notin P_i$ , and  $\tilde{S}_{ij} = S_{ij}$  otherwise, following the observation of Moreira et al. (2024) that mined hard negatives contain a substantial fraction of false negatives. Since mining retains only anchors with all four criteria populated (§3.2), no slot is ever filled at random.

### B.2 Objectives

**InfoNCE** The baseline uses polarity only: one positive, and every other slot in the pool a negative of equal status.

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^B -\log \frac{\exp(\tilde{S}_{i,p_i})}{\sum_{j=1}^M \exp(\tilde{S}_{ij})}$$

This is the standard formulation (Oord et al., 2018) in the instantiation of hard-negatives with one mine plus batch-negatives of DPR (Karpukhin et al., 2020).

**Weighted hard negative** Drops the mined easy negative and instead up-weights the anchor’s own hard negative by  $\alpha$ , implemented by adding  $\log \alpha$  to that logit before the softmax.

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^B -\log \frac{\exp(\tilde{S}_{i,p_i})}{\sum_{j \neq h_i} \exp(\tilde{S}_{ij}) + \alpha \exp(\tilde{S}_{i,h_i})}$$

where  $h_i$  indexes that slot. Hard negatives contributed by other anchors enter unweighted: their relation to query  $i$  has never been judged, so up-weighting them would amplify the false negatives the masking step removes. We use  $\alpha = 2.0$  which counts the mined hard negative twice in the denominator. This is the weighted hard-negative variant of supervised SimCSE (Gao et al., 2021), whose one positive and hard negative structure follows from NLI data, where a premise admits an entailment and a contradiction but no third unrelated sentence (Gao et al., 2021), we retain that structure rather than adding our mined easy negative to it. This configuration therefore differs from the InfoNCE baseline in three respects at once the mined easy negative is absent rather just  $\alpha$  addition.

**Multi-positive** Splits the positive side instead: both easy/hard cases are treated as positives, testing whether the hard positive carries signal rather than merely being spared from the negatives.

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^B \frac{-1}{|P_i|} \sum_{p \in P_i} \log \frac{\exp(\tilde{S}_{ip})}{\sum_{j=1}^M \exp(\tilde{S}_{ij})}$$

The average over positives sits outside the logarithm, so every positive must rank highly rather than only one of them, giving the  $\mathcal{L}_{out}^{sup}$  form of supervised contrastive loss (Khosla et al., 2020).<sup>10</sup>

**Graded triplet** Uses all four levels and unlike the above this approach don’t enable in-batch pool: every scored pair has been judged under our criteria. Scores are raw cosine similarities, since the margins are defined on that scale. With  $[\cdot]_+ = \max(0, \cdot)$

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^B \left( \begin{aligned} &[m_h - s_i^{ep} + s_i^{nh}]_+ \\ &+ [m_e - s_i^{ep} + s_i^{ne}]_+ \\ &+ [m_h - s_i^{hp} + s_i^{nh}]_+ \end{aligned} \right)$$

<sup>10</sup>This is not SupMPNRL (Dehghan and Amasyali, 2022), whose denominator excludes the anchor’s other positives and whose positives are copies of the anchor; ours are distinct mined documents at two graded levels.

| Configuration  | Data batch                         |
|----------------|------------------------------------|
| InfoNCE        | $c_{anom}, c_{nh}, c_{ne}$         |
| Weighted       | $c_{anom}, c_{nh}$                 |
| Multi-positive | $c_{anom}, c_{ph}, c_{nh}, c_{ne}$ |
| Graded triplet | $c_{anom}, c_{ph}, c_{nh}, c_{ne}$ |

Table 6: The training batch of anomaly variant,  $c_{anom}$  mean mined anomaly cases

where  $n$  is the number of active terms. Each term is the margin ranking loss (Schroff et al., 2015) in similarity form, equivalent under  $D = 1 - \cos$ . Setting  $m_e > m_h$  (0.30 and 0.10) is our extension: an unrelated case should end up further from the query than a hard negative that is superficially similar but legally distinguishable. The margins induce this ordering by capping the two negative levels at different distances below the positive; no term constrains  $s^{nh}$  against  $s^{ne}$  directly.

## C Qualitative Validation of Hard Negatives

For positive and anomaly mining, we require each candidate to share at least one charge with the anchor to satisfy the criteria described in Section 3.3. However during hard-negative mining, we do not filter candidates based on charge overlap. Instead, hard-negative selection relies the majority of LLMs. We find that LLMs consider some candidate cases as hard negative even they might share charges with the anchor. Figure 8 shows the charges most frequently shared by anchors and hard negatives in the test set. Most of these charges represent broad legal categories and therefore cannot, by themselves, establish that two cases concern the same legal situation. Table 4 illustrates this distinction using two drink-driving cases from the test set. In the anchor case, the defendant lost control of the vehicle, caused a fatal crash, and was charged with negligent homicide. In the hard-negative case, the defendant deliberately rammed a motorcycle and was therefore charged with attempted murder. Although both cases involve drink-driving, a majority of the LLM judges determine that the difference in intent and the resulting charges make the two cases legally distinct.

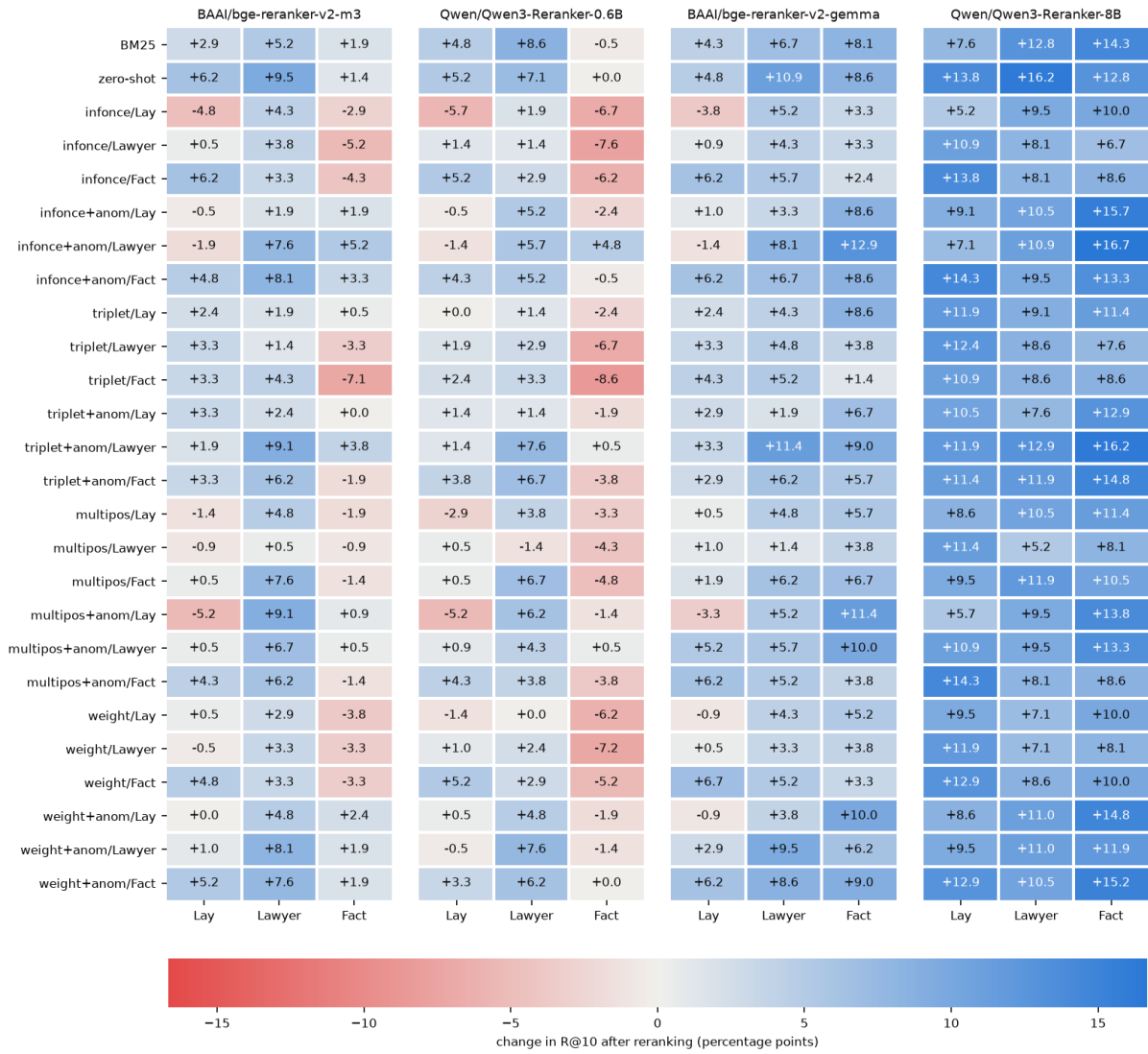


Figure 16: Heatmap of the performance difference ( $\Delta$ ) in R@10 before and after applying the reranker to the top-100 retrieved documents. Each row corresponds to a trained approach, in the format approach/training query type; each column is the test-set query type.

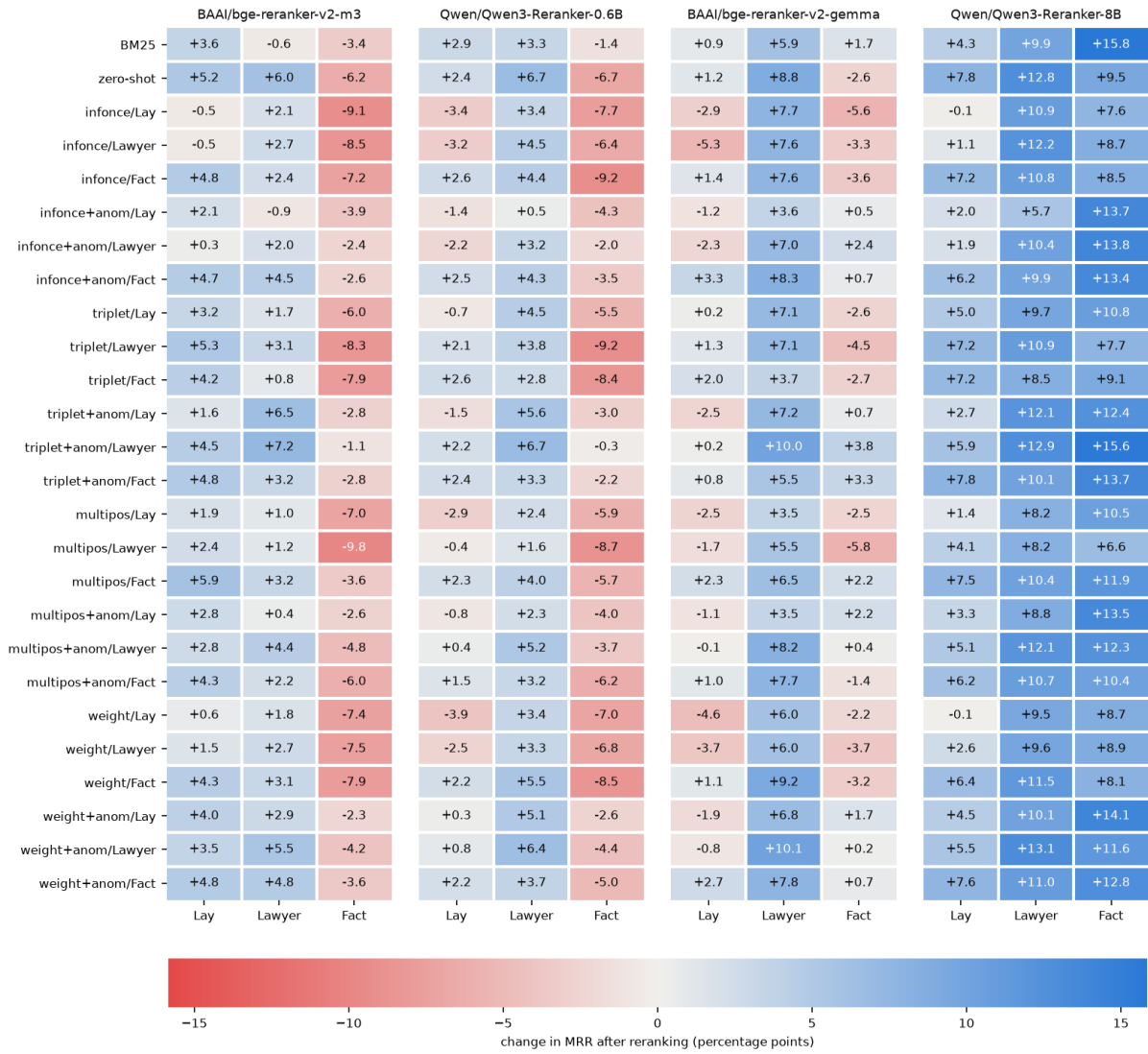


Figure 17: Heatmap of the performance difference ( $\Delta$ ) in MRR before and after applying the reranker to the top-100 retrieved documents. Each row corresponds to a trained approach, in the format approach/training query type; each column is the test-set query type.

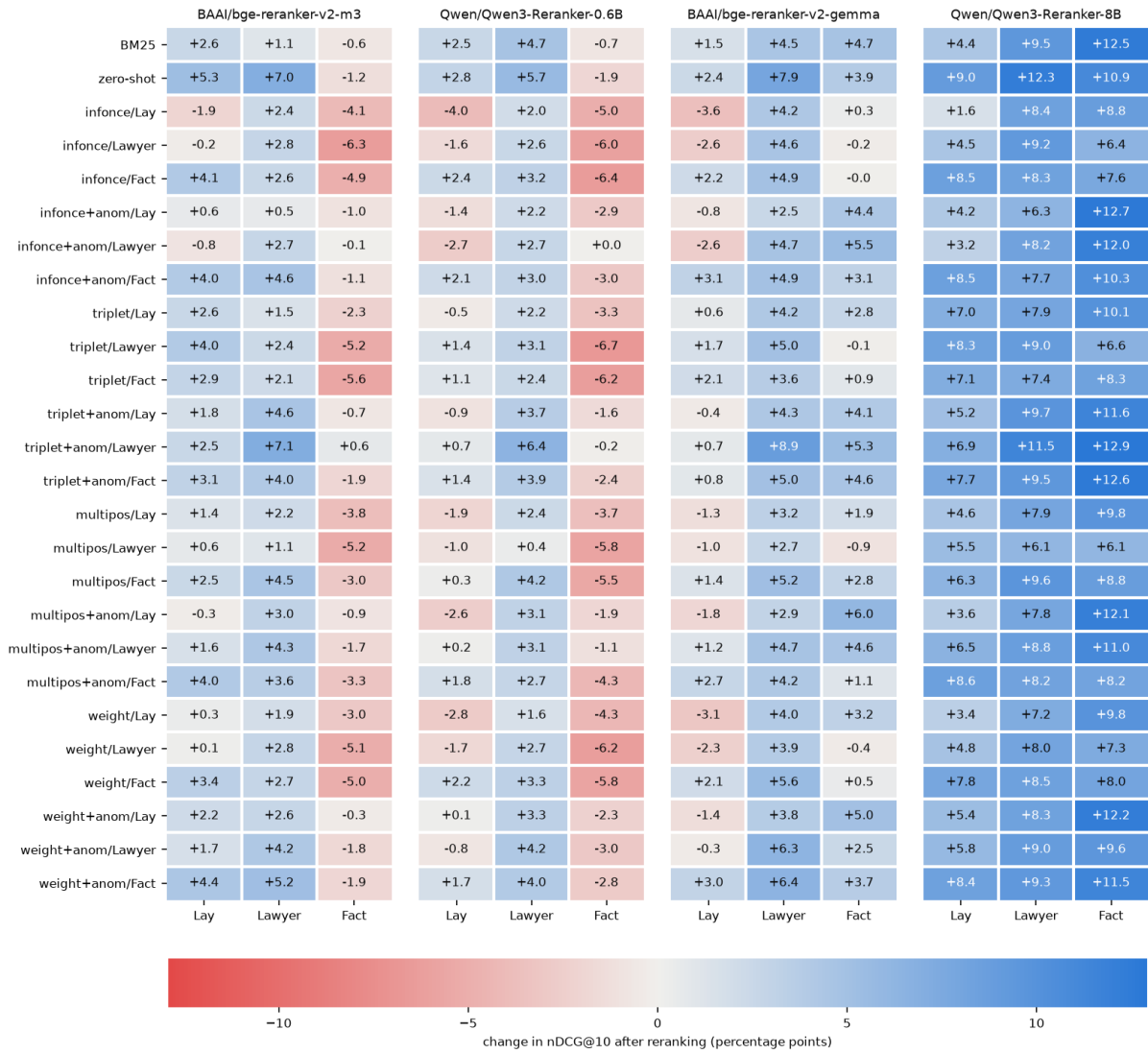


Figure 18: Heatmap of the performance difference ( $\Delta$ ) in nDCG@10 before and after applying the reranker to the top-100 retrieved documents. Each row corresponds to a trained approach, in the format approach/training query type; each column is the test-set query type.