SG-LegalCite: A Principle-Augmented Benchmark for Legal Citation Retrieval in Singapore Law

Shannon Lee Yueh Ern[1] **Kaidong Feng**[1] **Yingpeng Du**[2]
Chloe Lee En Jia[3] **Zhu Sun**[1]*
[1]Singapore University of Technology and Design, Singapore
[2]Nanyang Technological University, Singapore
[3]Yong Pung How School of Law, Singapore Management University, Singapore
yuehernshannon_lee@mymail.sutd.edu.sg, kaidong_feng@sutd.edu.sg, yingpeng.du@ntu.edu.sg
chloelee.2021@law.smu.edu.sg, zhu_sun@sutd.edu.sg

Abstract

Legal citation in common-law systems depends not only on factual similarity, but also on the legal principle for which a precedent is invoked. However, existing benchmarks for legal citation retrieval use case facts, citation context or full judgments as inputs, where the governing legal principle is often missing or only implicitly expressed and entangled with broader context. As a result, models may retrieve precedents that are factually similar yet doctrinally irrelevant. This limitation is particularly consequential in Singapore, where the legal system has evolved independently: only domestic precedents are binding, while foreign authorities serve merely as persuasive references. Thus, we propose a new retrieval paradigm that ranks cited cases based on queries integrating case facts and explicit legal principle, inspired by the real-world legal reasoning workflows. To support this paradigm, we further introduce SG-LegalCite, a dataset of 100,890 case–principle pairs extracted from 8,523 Singapore Supreme Court judgments spanning 2000–2025. Experiments across 11 baselines demonstrate the effectiveness of our principle-augmented retrieval paradigm, showing that explicit legal principles provide strong discriminative signals for legal citation retrieval. Our dataset and code is at https://github.com/ShannonLeeYuehErn/SG-LegalCite

1 Introduction

In common-law systems, legal citation is central to adjudication under *stare decisis*: courts justify outcomes by applying, distinguishing, or refining principles established in prior decisions (Woon, 1999). This has motivated growing interest in legal citation retrieval, which aims to identify precedents that support the reasoning of a citing judgment (Färber and Jatowt, 2020; Goebel et al., 2025). Typically, whether a precedent should be cited depends not only on factual similarity, but also on the legal principle for which it is invoked: cases with similar facts may implicate different doctrines and therefore support different lines of authority.

However, existing benchmarks for legal citation retrieval, as summarized in Table 1, always use case facts (Li et al., 2024, 2023b; Gao et al., 2024; T.y.s.s. et al., 2024), citation context (Hou et al., 2025; Wrzalik and Krechel, 2021; Mahari et al., 2024), or full judgments (Goebel et al., 2025; Mandal et al., 2017; Joshi et al., 2023) as queries, where the governing legal principle is often missing, only implicitly expressed, or entangled with broader context. Consequently, such systems may retrieve precedents that are factually similar yet doctrinally irrelevant. This reveals a fundamental limitation: existing benchmarks fail to explicitly model the doctrinal basis of legal citation.

This limitation is particularly consequential in Singapore (SG). Although rooted in the common-law tradition, its legal system has evolved independently: only domestic precedents are binding, while foreign authorities are merely persuasive. Thus, doctrinal alignment cannot be assumed even when factual patterns are similar. For instance, in negligence law, English courts apply the *Caparo* framework (House of Lords, 1990), whereas SG adopts the two-stage approach in *Spandeck* (Singapore Court of Appeal, 2007). A fact-based retrieval system may return precedents that are factually relevant but doctrinally inapposite. Despite this, no existing benchmark covers legal citation retrieval in SG.

To fill the gap, we propose a new retrieval paradigm where cited cases are ranked based on queries that integrate case facts with explicit legal principles, reflecting real-world legal reasoning workflows. To support such a paradigm, we introduce **SG-LegalCite**, a Singapore-specific dataset constructed from 8,523 Supreme Court judgments spanning 2000–2025. Through a multi-stage ex-

* Corresponding Author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1746–1766
November 6-10, 2026 ©2026 Association for Computational Linguistics

Paradigm	Dataset	Query	Output	Jurisdiction	Reference
Fact → Similar / Cited Cases	AILA	Raw fact from judgment	Similar cases	India	(Bhattacharya et al., 2019)
	LeCaRD	Raw fact from judgment	Similar cases	China	(Ma et al., 2021)
	MUSER	Raw fact + court opinion from judgment	Similar cases	China	(Li et al., 2023b)
	LeCaRDv2	Raw fact from judgment	Similar cases	China	(Li et al., 2024)
	LEAD	Fact summary via LLMs from judgment	Similar cases	China	(Gao et al., 2024)
	ECtHR-PCR	Raw fact from judgment	Cited cases	Europe	(T.y.s.s. et al., 2024)
	LEGAR	Raw fact from judgment	Similar cases	Korea	(Kim et al., 2025)
Citation Context → Cited Cases	GerDaLIR	Citation paragraph without cited cases	Cited cases	Germany	(Wrzalik and Krechel, 2021)
	LePaRD	Text before quotation (up to 300 words)	Quoted sentence	US	(Mahari et al., 2024)
	CLERC	Text before and after cited cases (150 words)	Cited cases	US	(Hou et al., 2025)
	AusLaw	Citation sentence + 1 preceding sentence	Cited cases	Australia	(Han et al., 2026)
Judgment → Cited Cases	FIRE IRLeD	Full judgment (cited cases masked)	Cited cases	India	(Mandal et al., 2017)
	IL-PCR	Full judgment (cited cases masked)	Cited cases	India	(Joshi et al., 2023)
	COLIEE	Full judgment (cited cases masked)	Cited cases	Canada	(Goebel et al., 2025)
	IL-PCSR	Full judgment (cited cases + statutes masked)	Cited cases + statutes	India	(Paul et al., 2025)
[Fact + Legal Principle] → Cited Cases	SG-LegalCite	Fact summary via LLMs from judgment + key legal principle	Cited cases	Singapore	–

Table 1: Legal retrieval benchmarks comparison.

traction pipeline, we obtain 100,890 case–principle pairs linking cited cases to the legal principles for which they are invoked. It enables controlled comparison between fact-only and principle-augmented retrieval, allowing direct evaluation of the role of legal principles. We further evaluate SG-LegalCite across 11 baselines. Results consistently show that principle-augmented retrieval outperforms methods that do not explicitly model legal principles, indicating that such principles provide strong discriminative signals for legal citation retrieval.

In summary, we establish a principle-augmented benchmark for legal citation retrieval in Singapore law. Our main contributions are threefold: (1) **A new retrieval paradigm**: we formulate legal citation retrieval as a principle-aware retrieval task that integrates case facts with legal principles; (2) **A new legal dataset**: we introduce SG-LegalCite, the first large-scale benchmark for legal citation retrieval in Singapore; (3) **Comprehensive evaluation**: we benchmark diverse retrieval methods and show that incorporating explicit legal principles substantially improves retrieval performance.

2 Related Work

Legal Retrieval Benchmarks. Legal citation retrieval systems typically retrieve precedents that best support the legal reasoning of a citing judgment, given a query derived from that case. Existing legal retrieval benchmarks can be categorized into three groups, as summarized in Table 1. One line of research adopts case facts as queries to retrieve similar or cited cases. For instance, AILA (Bhattacharya et al., 2019), LeCaRD (Ma et al., 2021), LeCaRDv2 (Li et al., 2024), and LEGAR BENCH (Kim et al., 2025) use raw facts from judgments to retrieve similar cases. Later, MUSER (Li et al., 2023b) introduces a multi-view query formulation (i.e., raw facts plus court opinions), while LEAD (Gao et al., 2024) leverages LLM-generated fact summaries from judgments instead of raw facts. Moreover, ECtHR-PCR (T.y.s.s. et al., 2024) adopts raw facts to retrieve cited cases, arguing that *fact-based queries better reflect the realistic pre-verdict setting*, since the court's reasoning is unavailable at the point of retrieval. LeCoPCR (T.y.s.s. et al., 2025) augments these facts with noun phrases for retrieval. Another line of work exploits citation contexts (e.g., the text surrounding a cited case) as queries to retrieve cited cases, as in GerDaLIR (Wrzalik and Krechel, 2021), CLERC (Hou et al., 2025), and AusLaw Citation Benchmark (Han et al., 2026), or to retrieve quoted sentences, as in LePaRD (Mahari et al., 2024). A third line of research directly uses the full judgment as the query to retrieve cited cases, as in COLIEE (Goebel et al., 2025), FIRE-IRLeD (Mandal et al., 2017), IL-PCR (Joshi et al., 2023), and IL-PCSR (Paul et al., 2025).

Overall, existing benchmarks have two key limitations. First, their queries–whether case facts, citation contexts, or full judgments–do not explicitly encode the legal principle for which a precedent is invoked. Second, they focus primarily on jurisdictions, such as Canada, India, China, Korea, Australia, Europe, and US, with no coverage of SG. Although SG belongs to the common-law tradition, its legal system has evolved independently: only domestic precedents are binding, while foreign authorities serve merely as persuasive references.

Legal Citation Recommendation. Citation recommendation aims to predict relevant references from a query context. In general domains, prior work includes probabilistic models (He et al., 2010), neural methods (Bhagavatula et al., 2018; Ebesu and Fang, 2017; Jeong et al., 2020), and graph-based meth-

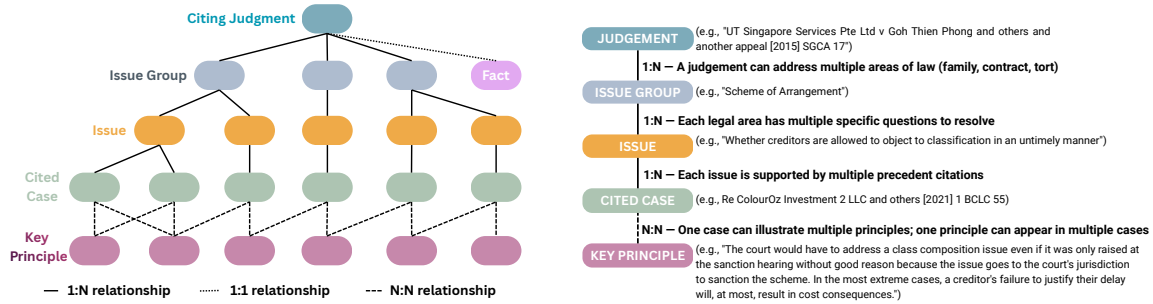


Figure 1: (Left) Data schema of SG-LegalCite; (Right) Example instances following the data schema.

ods that use citation structure (Cohan et al., 2020; Färber and Jatowt, 2020). However, in the legal domain, citation recommendation remains underexplored. Existing studies mostly predict legal citations from local textual context within judgments, including neural classifier methods (Huang et al., 2021), prototype-based methods (Luo et al., 2023), instruction-tuned models (Wang et al., 2024), and graph-based methods (Wendlinger et al., 2026). However, these methods differ from our setting in that they predict citations either from local textual context within judgments (e.g., citation contexts) or from co-citation relationships among similar citing judgments, rather than retrieving cases using queries that explicitly integrate case facts with legal principles.

3 The Proposed Retrieval Paradigm

Given a query q, we formulate legal citation retrieval as the task of retrieving relevant cited cases from a candidate pool $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$. Each query consists of the factual description of the citing judgment together with the legal principle under which authority is sought, denoted $q = [\textit{facts};\ \textit{principle}]$. The model generates a ranked list over $\mathcal{C}$, with the goal of ranking the ground-truth cited cases $\mathcal{C}^+ \subset \mathcal{C}$ as highly as possible. Formally, the task is to learn a scoring function $\mathcal{F}(q, c)$ such that $\mathcal{F}(q, c^+) > \mathcal{F}(q, c^-)$ for relevant cases $c^+ \in \mathcal{C}^+$ and irrelevant cases $c^- \in \mathcal{C} \setminus \mathcal{C}^+$. Our formulation differs from prior legal retrieval settings in that the query explicitly includes the legal principle associated with the citation, enabling retrieval to be conditioned not only on case facts but also on the doctrinal basis for which authority is invoked.

We define the principle as a legal construct rather than only operationally. The principle is the proposition of law for which the citing court invokes the cited case as authority (Cross and Harris, 1991; Garner et al., 2016; Duxbury, 2008), that is, the *citation-purpose proposition*. It may be composite, such as a multi-limb test or a set of related propositions, and it is stated as the citing court characterises it (Stone, 1959; MacCormick, 1987; Llewellyn, 1930). This construct is grounded in the doctrine of precedent, under which a case is cited as authority for a proposition; identifying that proposition is the function of citators. It is also what our extraction targets by design, since the citation context is precisely where the citing court states what the cited case stands for (See Section 4.3).

4 Dataset Construction

4.1 Data Schema

SG-LegalCite is constructed to support our proposed retrieval paradigm, where the query combines case facts with explicit legal principles. Accordingly, each record in SG-LegalCite contains five fields: the factual summary of the citing judgment (f), the legal principle for which a precedent is invoked (k), the cited case (c), together with a fine-grained doctrinal tag (*issue group*; e.g., "Damages", "Contract") and a specific legal issue (*issue*; e.g., "Whether the duty of care was breached"). Specifically, the triplet (f, k, c) directly instantiates our paradigm in Section 3: f and k form the query, and c is the retrieval target. The issue provides the legal question being resolved in the judgment, providing necessary context to identify the correct principle within potentially long citation passages and avoid peripheral statements. In addition, issue and issue group serve as metadata: issue group enables filtering and evaluation by doctrinal tag (e.g., Damages), while issue supports finer-grained analysis of model performance and error patterns across specific doctrinal questions.

Figure 1 shows the data schema: hierarchical relationships among judgments, fact, issue groups, issues, cited cases, and key principles. Each judgment contributes one shared facts field f and refers to multiple individual citations each associated with its own principle k, cited case c, issue and issue

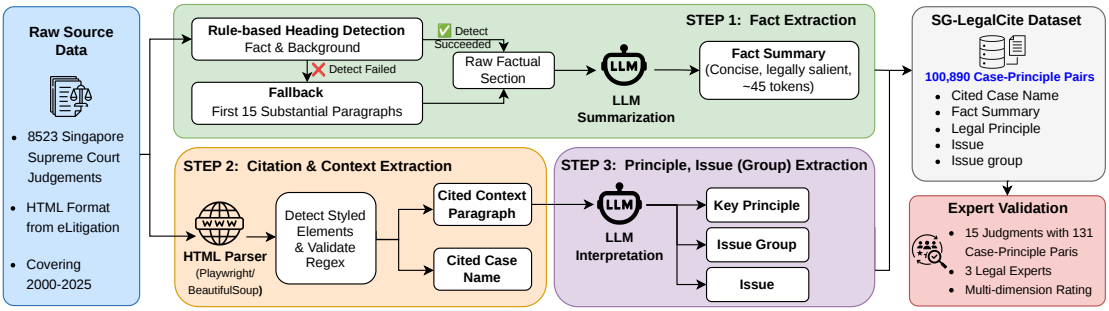


Figure 2: SG-LegalCite dataset construction pipeline.

group. The structure reflects the process of legal reasoning: the facts describe the dispute as a whole, while individual citations are invoked for legal principles relevant to particular aspects of that dispute.

4.2 Data Source

We collected 8,523 Supreme Court judgments spanning 2000–2025 from the eLitigation platform[1], the official public repository for Singapore court judgments. The corpus covers the Court of Appeal, Court of Appeal International, and High Court (including Registrar and Family divisions). Specifically, eLitigation's full archive contains 10,372 judgments (2000-2025) and we identified 9,404 valid judgment URLs. Of these, 8,523 judgments contained at least one cited case and were retained; 881 judgments contained zero cited cases and were dropped as they cannot contribute case–principle pairs to the retrieval task. Singapore judgments follow a broadly consistent structure: introduction and background sections describe the parties, procedural posture, and facts, while later issue-specific analysis sections contain the court's reasoning and citations to precedent. This regularity makes it feasible to separate factual background from citation-bearing doctrinal discussion, which is essential for constructing triplets of the form (f, k, c).

4.3 Extraction Pipeline

Since the corpus has 8,523 judgments, full manual annotation would require substantial expert effort (estimated at approximately **4-5 hours** per citing judgment) and is therefore impractical at this scale. We thus propose an automatic construction pipeline using large language model (LLM) (Liu et al., 2024) driven extraction with few-shot prompting, followed by expert validation on a sampled subset to assess quality. Figure 2 shows the pipeline, consisting of three main steps.

Step 1: Fact Extraction. For each judgment, we first locate the factual section through rule-based heading detection over the HTML structure. We prioritise headings such as *Facts*, *Background*, *Introduction*, and *Dispute*, and fall back to alternative factual headings where necessary. If no reliable heading is found, we extract the first 15 substantial paragraphs from the opening part of the judgment. Duplicated text is removed before summarization.

Raw factual sections are often lengthy and interspersed with procedural history and rhetorical detail that are irrelevant to the retrieval task. So, we truncate the text to 3,000 words and then summarize it with LLMs (e.g., Deepseek) via a structured prompt[2]. The LLM is asked to produce a concise 2–3 sentence, lawyer-style summary that preserves legally salient facts while removing procedural and rhetorical detail. This step reduces the average fact length from 1,034 tokens to 45 tokens, yielding compact factual descriptions suitable for retrieval. This compression level is designed to reflect how legal practitioners formulate retrieval queries in practice, that is, a concise and legally salient summary rather than the full factual background; it is also necessary to fit input limits of several encoder baselines (e.g., Legal-BERT, SBERT). An audit of the fact summaries for citation identifiers and doctrinal leakage is reported in Appendix L.

Step 2: Citation and Context Extraction. We use Playwright[3] for web scraping and BeautifulSoup[4] for HTML parsing. Cited case names are identified through styled HTML elements (e.g., `<em>`, `<i>`), with citation formats validated against Singapore neutral citation patterns (e.g., [2024] SGHC 123) and common foreign patterns. For each detected citation, we retain the surrounding ± 5 paragraphs as its local context.

[1] https://www.elitigation.sg

[2] All the prompts utilized in the data construction process are provided in Appendix A and Appendix B.

[3] https://playwright.dev

[4] https://beautiful-soup-4.readthedocs.io

Model	HSS	Cost per Case
Claude Sonnet 4 (Anthropic, 2025)	**91.3%**	$0.24
DeepSeek-V3 (Liu et al., 2024)	86.7%	**$0.02**
GPT-4o (OpenAI, 2024)	84.7%	$0.16

Table 2: Performance comparison of different LLMs on case-principle pair extraction task.

# Examples	HSS (Issue)	HSS (Key Principle)
0-shot	74.7%	86.7%
5-shot	76.7%	84.7%
10-shot	79.3%	85.3%
15-shot	**80.7%**	**86.7%**
20-shot	79.3%	86.7%

Table 3: Performance comparison across different numbers of in-context examples for issue and key-principle extraction using DeepSeek-V3.

Step 3: Principle, Issue and Issue Group Extraction. Legal principles are rarely stated as self-contained sentences. Instead, they are typically embedded in the court's reasoning alongside case comparisons, factual distinctions, and policy discussion. Extracting them thus requires interpreting the argumentative structure of the citation context, which is difficult to capture with rule-based or pattern-matching methods. We thus use an LLM to extract three fields from each citation paragraph extracted from step 2: (1) **Key Principles Illustrated** – the legal principle in full formal phrasing as stated in the judgment; (2) **Issue Group** – fine-grained doctrinal tag (e.g., "Damages", "Contract"); and (3) **Issue** – precise legal question (e.g., "Whether the duty of care was breached"). The LLM is instructed to preserve explicit legal formulations when present in the source text, with prompt examples spanning diverse legal domains (e.g., matrimonial assets, insolvency and statutory interpretation) to ensure robust extraction across case types. The extracted legal principle is then used as the principle field k in our retrieval task.

4.4 LLM Selection

To select the extraction model, we compare three LLMs (i.e., Claude Sonnet 4, DeepSeek-V3, and GPT-4o) against 25 judgments containing 725 case-principle pairs annotated by legal experts. We evaluate model outputs against these annotations using a *Hybrid Similarity Score (HSS)*, defined as the equally weighted average of ROUGE (Lin, 2004), which captures lexical overlap in legal terminology, and BERTScore (Zhang et al., 2020), which captures semantic similarity. Through this evaluation, we aim to determine which LLM provides the best accuracy–cost trade-off and how many in-context examples should be included in the prompt.

Attribute	Value
Time Span	2000–2025
Unique Judgments	8,523
Case-Principle Pairs	100,890
Unique Principles	72,500
Unique Cited Cases	48,478
Unique Issues	86,519
Unique Issue Groups	9,748
Avg. Raw Fact Length	1,034.4 tokens
Avg. Fact Length (post-summary)	45.1 tokens
Avg. Citation Paragraph Length	1,100.5 tokens
Avg. Principle Length	69.9 tokens

Table 4: SG-LegalCite dataset statistics.

We first compare all three LLMs under the same prompt setting. In Table 2, Claude achieves the highest accuracy (91.3%) at $0.24 per case, whereas DeepSeek attains 86.7% at $0.02 per case – a 12× cost reduction for only a 4.6% drop in HSS. DeepSeek thus offers the most favourable accuracy–cost trade-off among the three LLMs. We then further tune the number of in-context examples on DeepSeek. Table 3 shows that key principle extraction already performs strongly in the zero-shot setting (86.7%), with little variation across shot counts, consistent with legal principles often being stated relatively explicitly in judicial reasoning. Issue extraction, by contrast, benefits more from additional examples, improving from 74.7% in the zero-shot setting to 80.7% at 15-shot. So, we select DeepSeek with 15 examples for full-scale extraction, yielding a total cost of $78.22.

4.5 The Dataset and Expert Validation

Constructed Dataset. Table 4 shows the statistics of the resulting dataset. To characterise the doctrinal composition of the corpus, we classified all 8,523 judgments against the 34 practice-area tags used by Singapore Law Watch (Singapore Academy of Law, 2026). The most represented primary domains are Criminal Law (20.7%), Business & Commerce (10.9%), Civil Law & Procedure (8.5%), Tort (6.8%), and Family Law (6.8%); the full distribution is provided in Appendix E. We did not artificially balance the corpus across domains. SG-LegalCite is intended as a benchmark for retrieval systems deployed by Singapore legal practitioners. The natural distribution, dominated by criminal and commercial matters, with sparse coverage of emerging areas such as data protection and health care & life sciences, is itself the population a deployed system would face. The observed skew reflects the institutional role of the Supreme Court (apex criminal jurisdiction; commercial-hub case mix) and is consistent with how the Singapore Law Reports themselves catalogue judgments.

Citing Judgment	Issue
Re Ariffin Iskandar Sha bin Ali Akbar and other matters [2025] SGHC 156	Whether the stakeholders agree that Mr Foo should be admitted to the Bar.
Raw Fact (scraped Background section, 1,139 words, abbreviated)	**Citation Paragraph** (987 words, abbreviated)
Where the court determines that an applicant is not yet suitable for admission to the Bar on account of some issue of character, the usual course has been to invite the applicant to withdraw his or her application [...] The applications before me are three such Legacy Cases, which were due for consideration just as the Legal Profession (Admission) Rules 2024 took effect [...] while the applicants might not yet be fit for admission, withdrawal would not be appropriate.	The second incident occurred while Mr Foo was taking the LAW204 Constitutional & Administrative Law module [...] However, his essay contained phrases that appeared to have been lifted from Wikipedia without attribution [...] The AG, SILE and LSS accepted that the LAW204 Incident did not disclose dishonesty, but a lack of academic diligence: *Re Suria Shaik Aziz* [2023] 5 SLR 1272 at [25]. They therefore submitted that the LAW204 Incident did not affect Mr Foo's suitability of character. [...] Mr Foo did not contest the position taken by the AG and the SILE in respect of deferring his application.
Fact (LLM-summary)	**Key Principle Illustrated**
My client was found not yet fit for admission to the Bar due to a character issue, but his application was stayed for 18 months instead of dismissed because he qualified under the old admission rules. The stay allows him time for rehabilitation without forcing him to retake exams and restart his training period.	Dishonesty is to be distinguished from a lack of academic diligence.
Issue Group	**Cited Case**
Admission of Candidate	*Re Suria Shaik Aziz* [2023] 5 SLR 1272

Table 5: An example of SG-LegalCite extraction.

Table 5 provides a concrete example of a citation-level record produced by our pipeline, illustrating how raw factual and citation context are transformed into the structured fields in SG-LegalCite. Moreover, the breakdown of citation targets into domestic and foreign authorities, together with retrieval performance stratified by target type, is reported in Appendix H.

Quality Assessment. Expert validation is conducted on a randomly sampled subset to assess the extraction quality. Specifically, three legal experts (details in Appendix D) independently rated 131 case-principle pairs from 15 randomly sampled judgments using a 5-point scale across various dimensions on the key principle illustrated, issue and issue group (details in Appendix D). The overall mean ratings are 3.22 ($\sigma = 1.10$) for Key Principles Illustrated, 3.93 ($\sigma = 0.98$) for Issue, and 4.28 ($\sigma = 0.99$) for Issue Group, with $ICC(2, k) = 0.54$, indicating moderate inter-rater agreement. We regard this as consistent with the interpretive nature of legal principle formulation: a single citation may admit multiple defensible doctrinal formulations even when annotators broadly agree on the underlying legal point. The full annotation instrument, including the rating scale and the field-specific criteria given to annotators, is reproduced in Appendix M.

Comparison Across LLMs. To further validate the LLM selection in Section 4.4, the same three annotators independently rated 131 case–principle pairs extracted by Claude Sonnet 4 from the same 15 judgments under identical 1–5 scoring criteria, without being informed which model produced each set of extractions. Results show Claude achieving higher quality scores than DeepSeek: Key Principles Illustrated improved from 3.22 to 3.65 (+13.4%), Issue extraction from 3.93 to 4.35 (+10.7%), and Issue Group from 4.28 to 4.49 (+4.9%). Despite the quality improvements, DeepSeek-V3's substantially lower API cost makes it more practical for large-scale extraction (see Table 2). We therefore retain DeepSeek-V3 as the extraction model for SG-LegalCite.

5 Experiments

We conduct experiments on our SG-LegalCite to evaluate our proposed retrieval paradigm.

5.1 Experiment Setup

Baselines. We compare 11 methods in three categories: *conventional lexical models*, *conventional pre-trained language models* and *large-scale legal language models*. The **first category** includes one classic retrieval model BM25 (Rosa et al., 2021). The **second category** includes one general-purpose baseline – SBERT (Reimers and Gurevych, 2019), and six legal-specific models: Legal-BERT (Chalkidis et al., 2020), Custom Legal-BERT (Zheng et al., 2021), Pile-of-Law BERT(Henderson et al., 2022), SAILER (Li et al., 2023a), Legal-Longformer (Chalkidis et al., 2023), and Legal-en-RoBERTa (Niklaus et al., 2024). The **third category** has three models tailored for the legal domain: AdaptLLM (Cheng et al., 2024), SaulLM-7B (Colombo et al., 2024b), and Lawma-8B (Dominguez-Olmedo et al., 2025). This diverse set of models spans variations in parameter scale, pre-training corpus size, and pre-training objectives, thus enabling a comprehensive evaluation[5].

Evaluation Protocols. Following SOTA retrieval methods (Hou et al., 2025), we use Recall@k

[5]We did not compare with Lawyer GPT (Yao et al., 2024) and InternLM-Law (Fei et al., 2025) due to unavailability of released models; and did not compare with SaulLM-54B (Colombo et al., 2024a) due to GPU memory limits.

Model	Fact-only Query					Principle-augmented Query				
	MRR	R@1	R@5	R@10	R@20	MRR	R@1	R@5	R@10	R@20
Conventional Lexical Baseline										
BM25	1.8	0.6	2.1	3.2	4.9	$3.2^{+79\%}$	$2.0^{+221\%}$	$3.5^{+65\%}$	$4.7^{+45\%}$	$6.6^{+35\%}$
Conventional Pre-trained Language Models										
SBERT	10.5	4.7	15.1	22.5	31.2	$20.9^{+99\%}$	$12.9^{+174\%}$	$27.8^{+84\%}$	$36.3^{+61\%}$	$45.8^{+47\%}$
Legal-BERT	6.2	2.2	8.0	13.5	21.1	$14.1^{+128\%}$	$7.1^{+228\%}$	$19.6^{+145\%}$	$27.5^{+104\%}$	$37.5^{+78\%}$
Custom Legal-BERT	6.1	2.1	8.0	13.7	20.8	$9.2^{+51\%}$	$4.0^{+90\%}$	$12.4^{+55\%}$	$18.9^{+38\%}$	$27.4^{+32\%}$
Legal-Longformer	6.8	2.6	9.0	14.4	22.2	$10.0^{+47\%}$	$4.4^{+69\%}$	$13.9^{+54\%}$	$20.9^{+45\%}$	$29.9^{+35\%}$
Pile-of-Law BERT	5.7	2.0	7.3	12.2	18.9	$12.4^{+118\%}$	$5.8^{+190\%}$	$17.5^{+140\%}$	$25.5^{+109\%}$	$35.2^{+86\%}$
SAILER	0.8	0.1	0.5	1.0	2.1	$0.9^{+13\%}$	$0.2^{+100\%}$	$0.6^{+20\%}$	$1.2^{+20\%}$	$2.0^{-5\%}$
Legal-en-RoBERTa	6.9	2.6	9.1	14.6	21.9	$9.4^{+36\%}$	$3.8^{+46\%}$	$13.0^{+43\%}$	$20.0^{+37\%}$	$29.6^{+35\%}$
Large-scale Legal Language Models										
AdaptLLM	11.6	5.3	16.4	24.7	33.8	$29.5^{+154\%}$	$17.9^{+238\%}$	$41.6^{+154\%}$	$53.5^{+117\%}$	$64.9^{+92\%}$
Lawma-8B	10.0	4.0	14.1	21.7	31.7	$30.5^{+205\%}$	$18.2^{+355\%}$	$43.8^{+211\%}$	$56.6^{+161\%}$	$68.6^{+116\%}$
SaulLM-7B	**13.0**	**5.9**	**18.7**	**27.3**	**37.5**	$\mathbf{38.2^{+194\%}}$	$\mathbf{24.4^{+314\%}}$	$\mathbf{54.2^{+190\%}}$	$\mathbf{66.7^{+144\%}}$	$\mathbf{77.2^{+106\%}}$

Table 6: The Performance (%) of 11 baselines on our SG-LegalCite Dataset. Relative gains (shown as percentages in the upper-right corner of the MRR or Recall values for the principle-augmented query) are computed against the corresponding fact-only setting for the same model. The best results for each paradigm are highlighted in bold.

(R@k) and Mean Reciprocal Rank (MRR) as evaluation metrics, capturing both citation retrieval accuracy and ranking quality. k is set to {1, 5, 10, 20}. Results are reported from a single run per model due to the high computational cost of fine-tuning 11 models across different query settings. To ensure reproducibility, we fix the random seed to 42.

Query Settings. We fine-tune the 11 selected approaches on our SG-LegalCite dataset under two query settings: ***fact-only*** and ***principle-augmented***. In the fact-only setting, the retriever receives only the factual background of the citing judgment as the query. In the principle-augmented setting, the query concatenates the judgment facts with the corresponding legal principle, exposing both the factual pattern and the doctrinal basis of the citation. To speed up the testing process, we constructed a candidate pool: all models are tested on a 1000-way candidate pool consisting of one gold cited case and 999 randomly sampled negatives.

Implementation Details. For fine-tuning, we use contrastive learning with symmetric InfoNCE loss (van den Oord et al., 2018) over query–case pairs. The dataset is split into train, validation, and test sets with the 8:1:1 ratio. All models are trained with batch size 64 and learning rate 2e-5. The temperature parameter in InfoNCE loss is initialized at 0.07 and learned during training, and optimal hyperparameter settings are based on validation loss. All our experiments are conducted on NVIDIA A100 GPUs with 80 GB memory.

5.2 Main Results

Table 6 reports the performance of all models fine-tuned on our SG-LegalCite under the two query settings: *fact-only* and *principle-augmented*. Several major findings are noted.

(1) ***Large-scale legal language models consistently outperform smaller models across both query settings***. SaulLM-7B achieves the best performance, followed by SBERT, whereas BM25 performs the worst. This suggests legal citation retrieval benefits substantially from both model scale and domain-specific pre-training, enabling better modeling of relationships among case facts, legal principles, and precedents.

(2) ***Principle-augmented retrieval consistently outperforms fact-only retrieval***. Across nearly all models, augmenting queries with legal principles yields substantial gains. On average, it improves MRR by 111% and Recall by 117%, suggesting that explicit legal principles provide additional semantics that facilitate accurate cited-case retrieval.

(3) ***Large-scale legal language models benefit most from principle augmentation***. On average, our paradigm improves MRR by 79%, 70%, and 184%, and Recall by 92%, 77%, and 183% across the three model categories. The widened performance gap suggests stronger models better exploit principle-level semantics beyond lexical or sentence-level matching.

(4) ***Among conventional pre-trained language models, the general-purpose SBERT unexpectedly outperforms all legal-specific models***. It may be attributed to SBERT's retrieval-aligned contrastive pretraining, while most legal encoders rely

on masked language modeling. Besides, differences in corpus language, scale, recency, and jurisdictional coverage may limit the transferability of legal-specific models to Singapore law. Further studies are needed to isolate the contribution of each factor and we leave them as our future work.

(5) **Recall@k improvements show a consistent trend across models**. Gains are largest at Recall@1 and gradually decrease as k increases (184% $\rightarrow$ 106% $\rightarrow$ 80% $\rightarrow$ 60% for $k = 1, 5, 10, 20$ on average). The strongest improvements at top ranks indicate that principle-augmented queries are particularly effective at ranking the correct cited case first.

Query Setting	MRR	R@1	R@5	R@10	R@20
SBERT (all-mpnet-base-v2)					
Fact-only	10.5	4.7	15.1	22.5	31.2
Fact + Issue Group	19.4 $^{+85\%}$	10.7 $^{+127\%}$	27.5 $^{+82\%}$	37.4 $^{+66\%}$	49.0 $^{+57\%}$
Fact + Issue	20.8 $^{+98\%}$	11.5 $^{+145\%}$	29.6 $^{+96\%}$	39.7 $^{+76\%}$	50.2 $^{+61\%}$
Fact + Principle	20.9 $^{+99\%}$	12.9 $^{+174\%}$	27.8 $^{+84\%}$	36.3 $^{+61\%}$	45.8 $^{+47\%}$
Fact + All	**26.3** $^{+151\%}$	**16.5** $^{+251\%}$	**36.6** $^{+142\%}$	**46.1** $^{+105\%}$	**55.8** $^{+79\%}$
Fact + Context	18.1 $^{+72\%}$	9.4 $^{+100\%}$	26.3 $^{+74\%}$	35.8 $^{+59\%}$	46.7 $^{+50\%}$
SaulLM-7B (Equall/Saul-Instruct-v1)					
Fact-only	13.0	5.9	18.7	27.3	37.5
Fact + Issue Group	17.5 $^{+35\%}$	8.8 $^{+50\%}$	25.3 $^{+35\%}$	35.5 $^{+30\%}$	47.1 $^{+26\%}$
Fact + Issue	36.3 $^{+179\%}$	22.1 $^{+275\%}$	52.8 $^{+182\%}$	65.9 $^{+141\%}$	77.1 $^{+106\%}$
Fact + Principle	**38.2** $^{+194\%}$	**24.4** $^{+314\%}$	**54.2** $^{+190\%}$	**66.7** $^{+144\%}$	**77.2** $^{+106\%}$
Fact + All	34.8 $^{+168\%}$	21.9 $^{+271\%}$	49.3 $^{+164\%}$	61.5 $^{+125\%}$	72.9 $^{+94\%}$
Fact + Context	18.9 $^{+45\%}$	9.6 $^{+63\%}$	27.4 $^{+47\%}$	38.1 $^{+40\%}$	49.9 $^{+33\%}$

Table 7: Detailed Ablation Study. Relative gains (shown as superscript percentages) are computed against the corresponding fact-only setting.

5.3 Detailed Ablation Study

Principle vs. Other Doctrinal Signals. Our main results in Section 5.2 establish that augmenting queries with the legal principle consistently improves retrieval performance. A natural follow-up question is whether this gain reflects the particular discriminative value of formally phrased legal principles, or whether any structured doctrinal signal extracted from the same citation context would yield comparable improvements. So, we compare four types of doctrinal augmentation derived from SG-LegalCite against the fact-only baseline: (i) Fact + Issue Group, augmenting with the fine-grained doctrinal tag; (ii) Fact + Issue, augmenting with the specific legal question; (iii) Fact + Principle, augmenting with the extracted legal principle; and (iv) Fact + All, concatenating all the three doctrinal signals. Detailed experimental settings are in Appendix F. We evaluate them on two representative models: SBERT and SaulLM-7B.

Four observations emerge from Table 7. First, ***every form of doctrinal augmentation improves over the fact-only baseline across both models***, validating the broader value of structured legal signals at any granularity. Second, ***the benefit of granularity is mediated by model capacity***: SaulLM-7B is highly sensitive to signal specificity (Issue Group: +35% MRR; Issue: +179%; Principle: +194%), while SBERT gains comparably from all three single-signal settings (+85–99% MRR), suggesting that capable models exploit fine-grained doctrinal precision that weaker encoders cannot. Third, ***"Fact + Principle" is the strongest single doctrinal signal for SaulLM-7B***, confirming that formally phrased legal principles carry discriminative information compared with other signals. Finally, ***the "Fact + All" configuration is best for SBERT but underperforms "Fact + Principle" for SaulLM-7B***, indicating that concatenating overlapping signals provides complementary cues for weaker encoders but dilutes the already-sufficient principle signal for stronger models.

Principle vs. Citation Context. To further verify the efficacy of our proposed paradigm, we compare the extracted principles against raw citation context at matched text budgets. Details on citation context extraction is in Appendix G. We evaluate two representative models (SBERT and SaulLM-7B) under the "Fact + Context" setting. Table 7 (last row of each model) shows the results. "Fact + Context" improves substantially over "Fact-only" retrieval (SBERT: +72% MRR; SaulLM-7B: +45% MRR), confirming that citation-proximal text carries useful retrieval signal. However, "Fact + Principle" consistently outperform "Fact + Context" at the same text budget: SBERT achieves +15% higher MRR and +37% higher R@1 with principles, while SaulLM-7B shows +102% higher MRR and +154% higher R@1. The advantage is especially pronounced for SaulLM-7B, suggesting that larger models are better able to exploit the distilled doctrinal content in the principle field. These results demonstrate that the principle extraction step captures genuine doctrinal signal beyond what raw citation context provides, and that the gains reported in Table 6 are not attributable solely to richer input text.

Gold Principle vs. Predicted Principle. Our main experiments use the gold legal principle extracted from the citing judgment, which represents an oracle setting where the doctrinal basis of the citation is known a priori. In reality, a lawyer may begin with only a factual description of the case and may not have identified the governing principle. To as-

Setting	MRR	R@1	R@5	R@10	R@20
Fact-only	15.7	7.5	22.9	33.5	45.0
Cold-start	17.2 $^{+10\%}$	8.6 $^{+15\%}$	26.1 $^{+14\%}$	35.3 $^{+5\%}$	47.1 $^{+5\%}$
Gold (oracle)	**33.6** $^{+114\%}$	**20.8** $^{+177\%}$	**47.7** $^{+108\%}$	**61.8** $^{+84\%}$	**73.9** $^{+64\%}$

Table 8: Gold principle vs. predicted principle. Cold-start evaluation of fine-tuned SaulLM-7B on 1,000 randomly sampled test pools; superscripts are relative gains over fact-only. All values are percentages (%).

sess whether our principle-augmented paradigm is robust to this cold-start scenario, we test the fine-tuned SaulLM-7B under our "Fact + Principle" paradigm across three query settings: (i) Fact-only using fact as the query; (ii) Cold-start using fact and the predicted principle; and (iii) Gold (oracle) using fact and the extracted principle in our constructed datasets. Specifically, we randomly sample 1,000 cases from the test set. For each, we use DeepSeek-V3 to predict the legal principle from the fact summary alone[6], with case names disallowed in the output to prevent leakage. During inference, only the query text differs while all other settings are kept the same to ensure a fair comparison.

As in Table 8, "Cold-start" substantially exceeds "Fact-only" (+10% MRR, +15% R@1), confirming that the principle-augmented paradigm provides real benefit even when principles are predicted from facts alone. Because the cold-start principle is predicted from the facts alone and never from the citation context that contains the cited case, this gain cannot be attributed to leakage from that context. The gap between "Cold-start" and "Gold" implies clear headroom for improvement in principle prediction, e.g., through fine-tuned principle predictors or lawyer-in-the-loop refinement. Two further evaluations, one using same-domain hard negatives and one using a multi-positive fact-only setting, are reported in Appendix I and Appendix K. Building on the hard-negative evaluation, Appendix J breaks the improvement from fact-only to gold-principle queries down by legal domain, under both random and same-domain pools.

6 Conclusion and Future Work

We introduce SG-LegalCite, a principle-augmented benchmark for legal citation retrieval. First, we propose a novel retrieval paradigm that integrates legal principles with case facts, moving beyond fact-only methods. Second, to support such a paradigm, we construct the first large-scale dataset for Singapore law, comprising 100,890 case–principle pairs from 8,523 Supreme Court judgments (2000–2025). Experiments show the effectiveness of our proposed paradigm, demonstrate that explicit legal principles provide strong discriminative signals for citation retrieval and underscore the importance of modeling both factual similarity and doctrinal relevance.

[6]The principle prediction prompt is in Appendix C.

Limitations

First, it relies on LLM-generated summaries and extracted principles with only partial expert validation, which may introduce biases and propagate upstream errors into retrieval performance. Second, experiments are conducted on a sampled candidate pool (1,000 cases) for inference efficiency, which may not fully reflect large-scale retrieval settings. Future work will explore more effective candidate selection methods and evaluation on the full candidate pool. Lastly, results are reported from a single run per model with a fixed random seed due to computational constraints, limiting the assessment of variance and robustness.

Acknowledgment

This research is supported by SMU-SUTD Internal Research Grant Call (SMU-SUTD 2023_02_01). LLMs were used to support coding and manuscript polishing. All scientific design, methodology, and analysis are the authors' own work.

References

Anthropic. 2025. Introducing claude 4. https://www.anthropic.com/news/claude-4.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 238–251. Association for Computational Linguistics.

Paheli Bhattacharya, Kripabandhu Ghosh, Saptarshi Ghosh, Arindam Pal, Parth Mehta, Arnab Bhattacharya, and Prasenjit Majumder. 2019. Overview of the FIRE 2019 AILA track: Artificial intelligence for legal assistance. In *FIRE (Working Notes)*, pages 1–12.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English multinational legal language model development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15535, Toronto, Canada. Association for Computational Linguistics.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. In *International Conference on Learning Representations*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282. Association for Computational Linguistics.

Pierre Colombo, Telmo Pires, Malik Boudiaf, Rui Melo, Dominic Culver, Etienne Malaboeuf, Gabriel Hautreux, Johanne Charpentier, and Michael Desa. 2024a. SaulLM-54B & SaulLM-141B: Scaling up domain adaptation for the legal domain. In *Advances in Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates, Inc.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, André F. T. Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and Michael Desa. 2024b. SaulLM-7B: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Rupert Cross and J. W. Harris. 1991. *Precedent in English Law*, 4th edition. Clarendon Law Series. Clarendon Press, Oxford.

Ricardo Dominguez-Olmedo, Vedant Nanda, Rediet Abebe, Stefan Bechtold, Christoph Engel, Jens Frankenreiter, Krishna P. Gummadi, Moritz Hardt, and Michael Livermore. 2025. Lawma: The power of specialization for legal annotation. In *International Conference on Learning Representations*.

Neil Duxbury. 2008. *The Nature and Authority of Precedent*. Cambridge University Press, Cambridge.

Travis Ebesu and Yi Fang. 2017. Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1093–1096, New York, NY, USA. ACM.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: Approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405.

Zhiwei Fei, Songyang Zhang, Xiaoyu Shen, Dawei Zhu, Xiao Wang, Jidong Ge, and Vincent Ng. 2025. InternLM-law: An open-sourced Chinese legal large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9376–9392. 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, Miami, Florida, USA. Association for Computational Linguistics.

Bryan A. Garner, Carlos Bea, Rebecca White Berch, Neil M. Gorsuch, Harris L. Hartz, Nathan L. Hecht, Brett M. Kavanaugh, Alex Kozinski, Sandra L. Lynch, William H. Pryor, Jr., Thomas M. Reavley, Jeffrey S. Sutton, and Diane P. Wood. 2016. *The Law of Judicial Precedent*. Thomson Reuters, St. Paul, MN.

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 20th International Conference on Artificial Intelligence and Law (ICAIL 2025)*, pages 506–515, Chicago, IL, USA. Association for Computing Machinery.

Jiuzhou Han, Paul Burgess, and Ehsan Shareghi. 2026. Legal citation prediction with LLMs: A comparative evaluation of instruction tuning, retrieval, and jurisdiction-specific pre-training on the AusLaw citation benchmark. *Artificial Intelligence and Law*.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and C. Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 421–430, New York, NY, USA. ACM.

Peter Henderson, Mark S Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel E Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. In *Advances in Neural Information Processing Systems*, volume 35, Red Hook, NY, USA. Curran Associates, Inc.

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 U.S. legal case retrieval and retrieval-augmented analysis generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7913–7928, Albuquerque, New Mexico. Association for Computational Linguistics.

House of Lords. 1990. Caparo Industries plc v Dickman. [1990] 2 AC 605.

Zihan Huang, Charles Low, Mengqiu Teng, Hongyi Zhang, Daniel E. Ho, Mark S. Krass, and Matthias

Grabmair. 2021. Context-aware legal citation recommendation using deep learning. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, pages 79–88, New York, NY, USA. ACM.

Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with BERT and graph convolutional networks. *Scientometrics*, 124(3):1907–1922.

Abhinav Joshi, Akshat Sharma, Sai Kiran Tanikella, and Ashutosh Modi. 2023. U-CREAT: Unsupervised case retrieval using events extrAcTion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Chaeeun 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, Suzhou, China. Association for Computational Linguistics.

Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023a. 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, New York, NY, USA. ACM.

Haitao Li, Yunqiu Shao, Yueyue Wu, Qingyao Ai, Yixiao Ma, and Yiqun Liu. 2024. 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. ACM.

Qingquan Li, Yiran Hu, Feng Yao, Chaojun Xiao, Zhiyuan Liu, Maosong Sun, and Weixing Shen. 2023b. MUSER: A multi-view similar case retrieval dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5336–5340. ACM.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Karl N. Llewellyn. 1930. *The Bramble Bush: Some Lectures on Law and Its Study*. Columbia University School of Law, New York.

Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2023. Prototype-based interpretability for legal citation prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4883–4898. Association for Computational Linguistics.

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. ACM.

Neil MacCormick. 1987. Why cases have rationes and what these are. In Laurence Goldstein, editor, *Precedent in Law*, pages 155–182. Clarendon Press, Oxford.

Robert Mahari, Dominik Stammbach, Elliott Ash, and Alex Pentland. 2024. LePaRD: A large-scale dataset of judicial citations to precedent. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9863–9877, Bangkok, Thailand. Association for Computational Linguistics.

Arpan Mandal, Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal, and Saptarshi Ghosh. 2017. Overview of the FIRE 2017 IRLeD track: Information retrieval from legal documents. In *FIRE (Working Notes)*, pages 63–68.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. MultiLegalPile: A 689GB multilingual legal corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

Shounak Paul, Dhananjay Ghumare, Pawan Goyal, Saptarshi Ghosh, and Ashutosh Modi. 2025. IL-PCSR: Legal corpus for prior case and statute retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14588–14611, Suzhou, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto Lotufo, and Rodrigo Nogueira. 2021. Yes, bm25 is a strong baseline for legal case retrieval. *arXiv preprint arXiv:2105.05686*.

Singapore Academy of Law. 2026. Singapore law watch. https://www.singaporelawwatch.sg/. Accessed: 14 September 2026.

Singapore Court of Appeal. 2007. Spandeck Engineering (S) Pte Ltd v Defence Science & Technology Agency. [2007] SGCA 37.

Julius Stone. 1959. The ratio of the ratio decidendi. *The Modern Law Review*, 22(6):597–620.

Santosh T.y.s.s., Rashid Haddad, and Matthias Grabmair. 2024. ECtHR-PCR: A dataset for precedent understanding and prior case retrieval in the European court of human rights. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 5473–5483, Torino, Italia. ELRA and ICCL.

Santosh T.y.s.s., Isaac Misael Olguín Nolasco, and Matthias Grabmair. 2025. LeCoPCR: Legal concept-guided prior case retrieval for European court of human rights cases. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1654–1661, Albuquerque, New Mexico. Association for Computational Linguistics.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Jie Wang, Kanha Bansal, Ioannis Arapakis, Xuri Ge, and Joemon M. Jose. 2024. Empowering legal citation recommendation via efficient instruction-tuning of pre-trained language models. In *Proceedings of the 46th European Conference on Information Retrieval*, pages 310–324. Springer.

Lorenz Wendlinger, Simon Alexander Nonn, Abdullah Al Zubaer, and Michael Granitzer. 2026. The missing link: Joint legal citation prediction using heterogeneous graph enrichment. In *Proceedings of the 36th International Conference on Database and Expert Systems Applications*, volume 16047 of *Lecture Notes in Computer Science*, pages 197–211. Springer.

Walter Woon. 1999. The doctrine of judicial precedent. In Kevin Y.L. Tan, editor, *The Singapore Legal System*, 2nd edition. Singapore University Press, Singapore.

Marco Wrzalik and Dirk Krechel. 2021. GerDaLIR: A German dataset for legal information retrieval. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 123–128, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shunyu Yao, Qingqing Ke, Qiwei Wang, Kangtong Li, and Jie Hu. 2024. Lawyer GPT: A legal large language model with enhanced domain knowledge and reasoning capabilities. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, pages 108–112. ACM.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? Assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law*, pages 159–168, New York, NY, USA. ACM.

A Fact Extraction Prompt and Examples

The following prompt was used for judgment-level fact extraction (Section 4). The model was instructed to produce a concise lawyer-style summary of the case facts, suitable for use as a retrieval query.

Task instruction: Rewrite the key facts of this case as a lawyer would describe their client's situation to a colleague — conversational, practical, and focused on what matters legally.

Rules:

- Write 2–3 sentences maximum.
- Use phrases like "my client", "the other party", "the company" instead of full names.
- Focus on: what happened and what is at stake.
- Do not include legal questions or ask anything.
- Write as if you are a lawyer explaining the situation, not summarising a judgment.
- Do not start with any preamble such as "Of course", "Sure", or "Based on the provided text".
- Start directly with the facts.
- If the text does not contain enough information, respond with: "Insufficient information to summarize facts."

In-context examples:

- My client signed a 2-year clause preventing them from working for any competitor after leaving the company.
- My client transferred money by mistake to the wrong company and there is no contractual relationship between them.
- My client is going through a divorce after 15 years together. Her husband earned more but she was a stay-at-home mum taking care of three kids.

- A new bubble tea shop opened down the road with a name and cup design that customers keep mixing up with my client's brand.
- My client's customers have been accidentally buying a competitor's product thinking it was ours.
- The shareholders want to sue a director for losses from a failed project.
- We have medical reports showing our client's mental state was significantly impaired at the time of the offence.
- The document production exercise will cost our client over $200,000 and it seems disproportionate to what is actually at stake.

B Principle, Issue and Issue Group Extraction Prompt and Examples

The following prompt was used for citation-level principle extraction (Section 4). The model was instructed to act as an expert legal analyst and extract three structured fields from each citation paragraph: *Key Principles Illustrated*, *Issue Group*, and *Issue*. Output was required in TSV format for direct downstream processing.

System prompt: *You are an expert legal analyst. Your job is to extract structured outputs for legal analytics.*

Task instruction: Given a paragraph from a judgment, extract exactly three fields: (1) *Key Principles* — use the full formal legal phrasing reflected in the paragraph; do not simplify unless removing procedural noise is strictly necessary; (2) *Issue Group* — the high-level legal category (e.g., "Division of matrimonial assets", "Damages", "Scheme of Arrangement"); (3) *Issue* — the precise legal issue in judgment style (prefer "Whether…", "When…", "Identifying…", or "Determining…").

Output format: A single TSV table with a header row and columns in the following order: `Cited Case`, `Paragraph`, `Key Principles Illustrated`, `Issue`, `Issue Group`. No additional keys, comments, or explanations outside the TSV. Full legal phrasing must be preserved; party names and citations are omitted unless essential to state the legal issue.

The 15 in-context examples used in the prompt are reproduced below.

Example 1

Input: The Husband argues that the court should draw an adverse inference against the Wife for failing to disclose bank statements of her account with DBS Bank Ltd. Adverse inferences are appropriate where there is a substratum of evidence that establishes a prima facie case against the person whom the inference is to be drawn, and that person must have had some particular access to the information he or she is said to be hiding: *UZN v UZM* [2021] 1 SLR 426 at [60]. I draw the adverse inference against the Wife that she had savings in the DBS Bank account. Adopting the uplift approach, I adjust the ratio for the division of the matrimonial assets by 5% in favour of the Husband.
Output:

- **Key Principles:** Adverse inferences are appropriate where there is a substratum of evidence that establishes a prima facie case against the person whom the inference is to be drawn, and that person must have had some particular access to the information he or she is said to be hiding.
- **Issue Group:** Division of matrimonial assets
- **Issue:** When adverse inferences should be drawn

Example 2

Input: ColourOz was decided after the release of the 2020 Practice Statement. There the English High Court reviewed the process for effecting a scheme of arrangement. Snowden J stated that: Whilst the court would always have to address a class question even if raised at sanction (because it goes to jurisdiction), the implicit warning now repeated in para 10 of the New Practice Statement is that unless a good reason can be shown, such a late submission is unlikely to be well received and might, in an extreme case, justify disallowing an opposing creditor's costs. In our view, the court would have to address a class composition issue even if it was only raised at the sanction hearing without good reason because the issue goes to the court's jurisdiction to sanction the scheme. Even in extreme cases, an objecting creditor's failure to justify their delay will, at most, result in cost consequences.
Output:

- **Key Principles:** The court would have to address a class composition issue even if it was only raised at the sanction hearing without good

reason because the issue goes to the court's jurisdiction to sanction the scheme. In the most extreme cases, a creditor's failure to justify their delay will, at most, result in cost consequences.

- **Issue Group:** Scheme of Arrangement
- **Issue:** Whether creditors are allowed to object to classification in an untimely manner

Example 3
Input: In *British Transport Commission v Gourley* [1956] AC 185 at 206, the House of Lords explained that in an action for personal injuries the damages are always divided into two main parts. First, there is special damage, which has to be specially pleaded and proved, consisting of out-of-pocket expenses and loss of earnings incurred down to the date of trial. Secondly, there is general damage which the law implies and is not specially pleaded, including compensation for pain and suffering and, if the injuries lead to continuing or permanent disability, compensation for loss of earning power in the future.
Output:

- **Key Principles:** For personal injury actions, damages are divided into two main parts. Special damages must be specially pleaded and proved, consisting of out-of-pocket expenses and loss of earnings incurred down to the date of trial. General damages are implied by law and include compensation for pain and suffering and the like.
- **Issue Group:** Damages
- **Issue:** Whether the heads of claim were general or special damages

Example 4
Input: A Review Committee is a "sifting mechanism" to weed out frivolous complaints operating within a tight statutory timeframe: s 85(6) and (8) of the LPA. Its function is to direct the Council to dismiss the matter if it is unanimously of the opinion that the complaint is "frivolous, vexatious, misconceived or lacking in substance". It is not the function of a Review Committee to go beyond the heads of complaint presented to it, and attempt to identify additional heads of complaint.
Output:

- **Key Principles:** A Review Committee is a 'sifting mechanism' to weed out frivolous complaints that operates within a tight statutory timeframe.
- **Issue Group:** Judicial Review
- **Issue:** Analysis of complaint and the Review Committee's report

Example 5
Input: To rely on the partial defence of diminished responsibility, an offender bears the burden of proving three cumulative requirements: (a) that he was suffering from an abnormality of mind; (b) that the abnormality of mind arose from a condition of arrested or retarded development of mind, any inherent cause, or was induced by disease or injury; and (c) the abnormality of mind substantially impaired his mental responsibility for his acts and omissions. "Abnormality of mind" refers to a state of mind so different from that of ordinary human beings that the reasonable man would term it abnormal.
Output:

- **Key Principles:** To rely on the partial defence of diminished responsibility, an offender bears the burden of proving three cumulative requirements: (a) suffering from an abnormality of mind; (b) that the abnormality arose from arrested development, inherent cause, or disease or injury; and (c) the abnormality substantially impaired his mental responsibility. Where the accused premeditates to kill under a veneer of rationality but the decision is the product of a disordered mind, the accused must show that but for the abnormality he would not have made that decision, and that in executing the intention he had no realistic moment of rationality enabling him to resile.
- **Issue Group:** Defence of diminished responsibility
- **Issue:** Whether the partial defence of diminished responsibility under Exception 7 to s 300 of the Penal Code is available to the offender

Example 6
Input: A plaintiff is entitled to plead and claim on the basis of both the wrongful gain interest and the wrongful loss interest in an action for breach of confidence. The law is clear that an action for breach of duty of confidence in equity and an action for breach of contractual duty of confidentiality are distinct causes of action and should not be conflated.
Output:

- **Key Principles:** A plaintiff is entitled to plead and claim on the basis of both the wrongful gain interest and the wrongful loss interest in an action for breach of confidence.
- **Issue Group:** Breach of duty of confidence
- **Issue:** Applicable law

Example 7
Input: In *Law Society of Singapore v Wong Sin Yee* [2018] 5 SLR 1261, the court found that by posing questions in cross-examination that were "indecent, scandalous and calculated to insult or annoy", the respondent solicitor's conduct was "disgraceful and a clear abuse of the privileges of cross-examination". Misconduct under the LPA is a broader catch-all provision where the conduct is deemed unacceptable but does not fall within any other enumerated ground. The standard applied is whether reasonable people, on hearing what the solicitor had done, would have said without hesitation that as a solicitor he should not have done it.
Output:

- **Key Principles:** By posing questions in cross-examination that were "indecent, scandalous and calculated to insult or annoy", the respondent solicitor's conduct was disgraceful and a clear abuse of the privileges of cross-examination. Misconduct under the LPA is a broader catch-all provision where the conduct does not fall within any other enumerated ground. The standard is whether reasonable people, on hearing what the solicitor had done, would have said without hesitation that as a solicitor he should not have done it.
- **Issue Group:** Disciplinary Proceedings
- **Issue:** Whether the respondent's conduct constitutes grossly improper conduct as a form of misconduct under the LPA

Example 8
Input: The English authorities make clear that the purpose of the statutory filter requiring authorisation for charity proceedings is to prevent the wasteful expenditure of precious charity funds on litigation. In *Dean v Burne* [2009] EWHC 1250 (Ch), Blackburne J took the view that s 33(4) did not obviate the need for authorisation where the counterclaim was in the nature of a wholly distinct claim and in no sense arose out of the subject matter of the action.
Output:

- **Key Principles:** Section 33(4) did not obviate the need for authorisation where the counterclaim was in the nature of a wholly distinct claim and in no sense arose out of the subject matter of the action.
- **Issue Group:** Statutory interpretation
- **Issue:** Statutory interpretation on authorisation

Example 9
Input: The implied undertaking arises from the principle established in *Riddick v Thames Board Mills Ltd* [1977] QB 881. Where a party to litigation has been ordered to give discovery, the discovering party may not use the discovered documents, and the information obtained therefrom, for a purpose other than pursuing the action in respect of which discovery is obtained. Public interest requires full and complete disclosure in the interest of justice, but discovery on compulsion is also an intrusion of privacy. The Riddick principle seeks to strike a balance between these two interests.
Output:

- **Key Principles:** Where a party to litigation has been ordered to give discovery, the discovering party may not use the discovered documents, and the information obtained therefrom, for a purpose other than pursuing the action in respect of which discovery is obtained. The Riddick principle seeks to strike a balance between the public interest in full disclosure and the intrusion of privacy caused by compelled discovery.
- **Issue Group:** Whether documents can be adduced as further evidence on appeal
- **Issue:** The Riddick principle

Example 10
Input: In *Yien Yieh Commercial Bank Ltd v Kwai Chung Cold Storage Co Ltd* [1989] 2 HKLR 639, Lord Goff stated that to reject one clause in a contract as inconsistent with another involves a rewriting of the contract which can only be justified where the two clauses are in truth irreconcilable. Where the document has been drafted as a coherent whole, repugnancy is extremely unlikely to occur. The contract has to be read as a whole, and the overwhelming probability is that an apparent inconsistency will be resolved by the ordinary processes of construction.
Output:

- **Key Principles:** To reject one clause in a contract as inconsistent with another involves a rewriting of the contract which can only be justified where the two clauses are in truth irreconcilable. Where the document has been drafted as a coherent whole, repugnancy is extremely unlikely to occur, and an apparent inconsistency will ordinarily be resolved by the ordinary processes of construction.
- **Issue Group:** Consistency between enforcement procedure and notice provisions
- **Issue:** Legal principles on whether the bank has a right to unwind

Example 11
Input: A quantitative assessment of the relative number of syllables which two marks have in common is usually conducted in ascertaining aural similarity. Allowances should be made for imperfect recollection and careless pronunciation. Ultimately, it is the pronunciation of the relevant words as a whole that must be critical in ascertaining aural similarity.
Output:

- **Key Principles:** It is the pronunciation of the relevant words as a whole that must be critical in ascertaining aural similarity.
- **Issue Group:** Trademark infringement
- **Issue:** Similarity between trademarks — aural similarity

Example 12
Input: The courts have indicated their willingness to imply a term of mutual trust and confidence where warranted. In *Johnson v Unisys Ltd* [2003] 1 AC 518, Lord Hoffmann observed that over the last 30 years the nature of the contract of employment has been transformed, and that a person's employment gives not only a livelihood but an occupation, an identity and a sense of self-esteem. The contribution of the common law has been by the evolution of implied terms in the contract of employment, the most far-reaching of which is the implied term of trust and confidence.
Output:

- **Key Principles:** The courts have indicated their willingness to imply a term of mutual trust and confidence where warranted. The most far-reaching contribution of the common law to employment contracts has been the implied term of trust and confidence.
- **Issue Group:** Bonuses
- **Issue:** Mutual trust and confidence to determine entitlement to bonuses

Example 13
Input: The elements of the tort of unlawful means conspiracy can be summarised as follows: (a) there must be a combination of two or more persons and an agreement between or among them to do certain acts; (b) the conspiracy must be done by "unlawful means", which covers both a criminal act as well as an intentional tortious act; (c) the act must actually be performed in furtherance of the agreement; (d) there must be an intention to injure the plaintiff; and (e) damage must be suffered by the plaintiff.
Output:

- **Key Principles:** The elements of the tort of unlawful means conspiracy include: (a) a combination of two or more persons and an agreement to do certain acts; (b) the conspiracy must be done by unlawful means; (c) the act must be performed in furtherance of the agreement; and (e) damage must be suffered by the plaintiff.
- **Issue Group:** Tort of unlawful means conspiracy
- **Issue:** Whether there is a claim under the tort of unlawful means conspiracy

Example 14
Input: In *Digicel (St Lucia) Ltd v Cable and Wireless plc* [2008] EWHC 2522 (Ch), Justice Morgan observed that what is generally required by an order for standard disclosure is a "reasonable search" for relevant documents, and that the rules do not require that no stone should be left unturned. As Lord Justice Jacob observed in *Nichia Corporation v Argos Limited* [2007] EWCA Civ 741: "Perfect justice" in one sense involves a tribunal examining every conceivable aspect of a dispute, but a system which sought such perfect justice in every case would actually defeat justice. A compromise is made: one makes do with a lesser procedure even though it may result in justice being rougher. Better justice is achieved by risking a little bit of injustice.
Output:

- **Key Principles:** Standard disclosure requires only a "reasonable search" for relevant documents. A system seeking perfect justice in every case would defeat justice due to disproportionate cost and time. Better justice is achieved

by accepting a lesser procedure, even at the risk of some roughness — better justice is achieved by risking a little bit of injustice.

- **Issue Group:** Electronic Discovery
- **Issue:** Extent to which inspection of compound documents is allowed

Example 15
Input: In *Re MC Bacon Ltd* [1990] BCLC 324, Millett J discussed the changes brought by the UK Insolvency Act 1986 in relation to unfair preference. It is no longer necessary to establish a dominant intention to prefer; it is sufficient that the decision was influenced by the requisite desire. It is no longer sufficient to establish an intention to prefer — there must be a desire to produce the effect mentioned in the subsection. It is the subjective desire of the company to improve the creditor's position in the event of its imminent insolvency that is decisive. There is no need for direct evidence of the requisite desire as it can be inferred from the circumstances of the case.
Output:

- **Key Principles:** It is no longer necessary to establish a dominant intention to prefer; it is sufficient that the decision was influenced by the requisite desire. There must be a desire to produce the effect mentioned in the subsection, not merely an intention to prefer.
- **Issue Group:** Unfair preference
- **Issue:** Whether the plaintiff was influenced by a desire to prefer the defendant

C Cold-Start Principle Prediction Prompt

The following prompt was used for cold-start principle prediction (Section 5.3). DeepSeek-V3 was instructed to predict the most likely legal principle from a fact summary alone, with case names disallowed in the output to prevent leakage of citation identity into the retrieval query. **System prompt:**

You are an expert legal analyst assisting a Singapore lawyer who needs to find precedent cases. The lawyer has described their client's situation but has not yet identified the specific legal authority they need.

Task instruction: Given a factual description of a case, predict the single most likely legal principle the lawyer would need precedent authority for. Output the principle in full formal legal phrasing as it might appear in a Singapore judgment.

Rules:

- Output exactly ONE legal principle.
- Use formal legal phrasing (e.g., "The duty of care arises when…", "An adverse inference may be drawn where…").
- Focus on the substantive legal proposition, not procedural rules unless procedure is the central issue.
- Do not name specific cases or cite case names.
- Do not include preamble such as "The principle is…" or "Based on these facts…".
- Output the principle directly as a self-contained statement.

Output format: A single sentence or short paragraph stating the legal principle. No additional text.

Input variable: `{fact_summary}` is replaced with the LLM-generated fact summary from Section 4 (Step 1).

Contamination safeguards. Before passing fact summaries to DeepSeek-V3, we screened them for Singapore neutral citation patterns (e.g., `[2024] SGHC 123`), "Party v Party" patterns, and citation keywords (e.g., SGCA, SGHC, SLR, MLJ); the same screen was applied to predicted principles. Across the 50 sampled test pools, no fact summary and no predicted principle was flagged.

D Data Consent, Recruitment And Payment

We engaged legal experts to annotate, e.g., case names, legal principles, issues, and issue groups based on citing judgments. All annotators were informed in advance that their contributions would be used solely for academic research purposes and provided informed consent prior to participation. We do not collect or disclose any sensitive personal information, including demographic details or private data of the annotators. All data handling procedures adhere to standard research ethics and privacy guidelines to ensure confidentiality and responsible use.

The three legal expert annotators were recruited in the expert validation process. Each annotator was compensated SGD 100 per grading round, with two rounds conducted in total — one for evaluating DeepSeek-V3 extractions and one for evaluating Claude Sonnet 4 extractions. The legal expert who

performed the initial annotations for LLM selection (25 judgments, 725 case-principle pairs) was compensated SGD 10 per judgment annotated.

Specifically, the three experts independently rated 131 case-principle pairs from 15 randomly sampled judgments using a 5-point scale across three dimensions on the key principle illustrated, issue and issue group.

- *Key Principles Illustrated*, assessed on whether the legal principles are accurately extracted from the cited paragraph, represent actual holdings or dicta, and are relevant to the citing context;
- *Issue*, assessed on whether it correctly identifies the legal question addressed, is sufficiently specific, and matches the citation context;
- *Issue Group*, a fine-grained doctrinal tag based on the issue, assessed on whether it accurately captures the doctrinal area of the citation and is consistent with the citation context.

E Domain Distribution

Figure 3 shows the distribution of the 8,523 judgments in SG-LegalCite across legal domains, classified using the 34 practice-area tags used by Singapore Law Watch (Singapore Academy of Law, 2026). Domains with fewer than 1.5% of the corpus are grouped as "Others (<1.5%)".

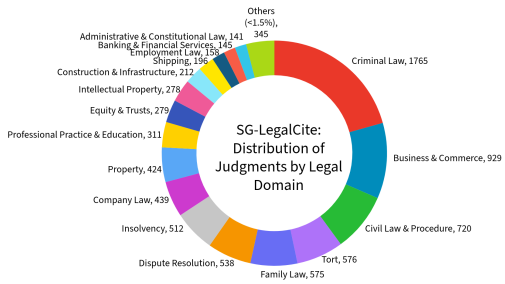


Figure 3: Distribution of judgments by legal domain in SG-LegalCite.

F Detailed Settings of Doctrinal Signal Granularity Ablation

For a fair comparison, all models share identical candidate pools, negative samples, and hyperparameters: batch size = 64; learning rate = 2e-5, 10 epochs with early stopping, QLoRA r=16 for SaulLM-7B; and only the query text differs.

G Detailed Settings of Citation Context Extraction

For each case–principle pair, we construct a *scrubbed context window*: a ±200-word window centred on the cited case mention in the original citation paragraph, from which all case names, neutral citations, reporter references, judge names, and pinpoint markers are removed to prevent retrieval via surface matching (residual identifiability <0.05% of cases). To ensure a fair comparison, the scrubbed context is then truncated to match the word count of the corresponding principle for that row ($\sim$ 70 tokens on average), so both query types receive the same text budget.

Regarding the residual identifiability, it is calculated in the following way. A multi-layer regex pipeline was applied to scrub the citation context windows. The pipeline removes: (1) the target cited case name; (2) all other case names in the paragraph, covering six patterns: Party v Party, PP v X, R v X, Re X, Ex parte X, and In the Matter of X; (3) Singapore neutral citations (e.g., [2024] SGCA 5); (4) reporter references (SLR, MLJ, AC, WLR, KB, etc.), including volume-reporter-page format (e.g., "1 KB 1"); (5) judge names with titles (e.g., Lord Denning, Phang JA, Stamp LJ); (6) bare year markers (e.g., [1985], (1848)); and (7) pinpoint references (e.g., at [60], at paras 14–15). Each match is replaced with a placeholder token ([CASE], [CITATION], [JUDGE], [YEAR], or [REPORTER]) to preserve sentence structure while removing identifiers. The scrubber was developed over four iterations (V1–V4), with a full-dataset audit after each round. V1 exhibited approximately 13% residual leakage, primarily from case names with single-letter parties (e.g., "R v S") and lowercase prefixes (e.g., "de Dampierre v de Dampierre") that the initial regex patterns did not capture. V4 addressed these by allowing single-character party names and lowercase prefix tokens (de, von, van, etc.) in the case-name regex. To verify the final scrubbing quality, we audited all 100,890 rows (not a sample) against 18 leakage-detection patterns. 13 of 18 patterns produced zero hits. The remaining five patterns yielded a combined total of approximately 50 rows with any surviving identifiable reference (22 real case-name leaks from unusual company-name formatting, plus 28 minor residuals such as bare "ex parte" phrases and stray year tokens), giving a residual identifiability rate of approximately 0.05%.

Domain	# Test Cases	Random: Fact-only → Gold	Hard: Fact-only → Gold
Criminal Law	226	16.5 → 34.9	7.0 → 18.0
Business & Commerce	150	9.8 → 33.4	6.1 → 21.6
Tort	85	14.6 → 26.3	5.7 → 14.8
Civil Law & Procedure	84	12.6 → 26.0	8.0 → 19.3
Insolvency	79	14.5 → 29.4	3.7 → 15.7
Dispute Resolution	74	13.0 → 26.9	2.3 → 7.8
Company Law	61	17.6 → 44.8	3.3 → 17.1
Family Law	52	23.1 → 41.7	3.8 → 11.0
Intellectual Property	49	20.9 → 33.7	5.2 → 9.8
Property	40	20.9 → 31.5	7.3 → 16.8
Professional Practice & Education	33	36.7 → 53.3	6.8 → 17.8
Administrative & Constitutional	23	15.3 → 28.4	3.7 → 11.7
Equity & Trusts	16	3.9 → 28.4	0.8 → 9.4
Employment Law	15	8.0 → 29.6	1.3 → 16.6
Banking & Financial Services	13	23.8 → 41.1	8.5 → 13.0

Table 9: MRR by legal domain under random and same-domain (hard) pools, fine-tuned SaulLM-7B.

H Domestic vs. Foreign Citation Targets

Of all citation instances in SG-LegalCite, 86.9% target Singapore cases and 13.1% target foreign authorities. Consistent with Singapore doctrine, foreign targets are annotated as *Persuasive* in 100% of instances (0 *Binding*), while domestic targets divide into *Binding* (17.2%), *Comity* (62.0%) and *Persuasive* (20.8%).

Query Setting	Domestic	Foreign
Fact-only	16.4	10.9
Cold-start	17.2	17.3
Gold	**32.8**	**38.4**

Table 10: MRR stratified by target type, on the 1,000-pool cold-start sample (fine-tuned SaulLM-7B).

Table 10 shows that foreign authorities are not easy targets. First, facts alone retrieve them poorly, and performance on foreign targets is substantially worse than on domestic ones. Second, gold principles achieve comparable performance on domestic and foreign targets, lifting MRR by +100% on domestic and +252% on foreign targets over fact-only. This is consistent with foreign cases being cited for doctrinal propositions rather than for factual similarity.

I Hard-Negative (Same-Domain) Evaluation

To test whether the random candidate pool rewards topic matching rather than doctrinal discrimination, we classified all 8,523 judgments into the 34 Singapore Law Watch practice-area domains and assigned each cited case a domain by majority vote over its citing judgments. We randomly sampled 1,000 test cases and rebuilt same-size pools in which all 999 negatives share the query's domain. Pool size (1,000), the queries, and the correct case are held identical to the random-pool condition, and the same fine-tuned SaulLM-7B ranks both.

Pool	Setting	MRR	R@1	R@5	R@20
Random	Fact-only	15.8	7.5	23.2	45.1
Random	Cold-start	17.7	8.2	27.0	49.5
Random	Gold	**33.3**	**19.3**	**49.3**	**73.5**
Hard	Fact-only	5.5	1.9	7.2	17.3
Hard	Cold-start	5.8	2.1	6.3	21.2
Hard	Gold	**16.2**	**7.7**	**24.0**	**43.5**

Table 11: Random vs. same-domain (hard) candidate pools, fine-tuned SaulLM-7B on 1,000 sampled test cases.

Same-domain negatives make the task considerably harder: fact-only MRR falls by 65%, confirming that random pools partly reward topic matching. The benefit of the principle, however, grows once that shortcut is removed: gold over fact-only rises from +111% MRR on random pools to +195% on same-domain pools.

J Retrieval Performance by Legal Domain

Using the evaluation of Appendix I, we break the 1,000 sampled test cases down by legal domain. Table 9 reports MRR (fact-only → gold) under both pool types.

Gold improves over fact-only in all 15 domains under both pool types (30/30 comparisons), so the aggregate result is not driven by the two largest domains. The relative benefit is largest in doctrinally driven domains, for example Insolvency (14.5 →

29.4, random) and Company Law (17.6 → 44.8, random), and smallest where fact patterns are most distinctive (Tort, Intellectual Property). This is consistent with facts already carrying partial signal in fact-differentiated disputes, while doctrinal domains depend on the principle. Domains with fewer than 20 test cases carry wide intervals and should be read with caution.

K Multi-Positive Fact-Only Evaluation

Our main evaluation grades each query against a single gold cited case. To quantify how much this understates the fact-only baseline, we re-ran the evaluation over the full test set of 850 judgments (10% of the 8,523 unique judgments), treating all cited cases of a test judgment as gold (mean 11.7, median 8) and randomly sampling negatives to fill each candidate set to 1,000.

Metric	@1	@5	@10	@20
Hit@k	37.1	67.8	76.8	84.4
Recall@k	5.1	18.1	28.5	41.7
nDCG	37.1	34.7	34.8	37.3
MAP		28.3		

Table 12: Multi-positive fact-only evaluation on the full test set, fine-tuned SaulLM-7B.

Single-positive grading does understate the fact-only baseline (Hit@5 = 67.8%, Table 12), although per-case recall remains low. We nonetheless adopt the single-positive setting throughout, for two reasons: it is applied identically to all compared methods, so the comparison remains fair; and under the Fact + Principle paradigm each query typically corresponds to a single cited case, which the single-positive pool reflects directly.

L Fact-Summary Leakage Audit

Fact summaries are extracted by section-heading matching (75%) or by a fallback method (25%). Because the fallback does not rely on an explicit *Facts* heading, we audited the summaries for material that could leak citation identity or doctrinal answers into the query.

A pattern audit detected no neutral citations and no citation keywords. Seven "X v Y" case names were found across all fact summaries (0.08%), and all seven are eponymous doctrine references (e.g. "the rule in *Rylands v Fletcher*") that cannot be removed without altering the substantive legal content of the summary. Issue statements and doctrinal vocabulary cannot be pattern-matched, as the space of possible phrasings is too large, so we used an LLM (DeepSeek-Chat) to judge a stratified sample of 308 summaries. Issue statements appear in 7.3% of heading-extracted summaries and 13.3% of fallback-extracted ones; doctrinal vocabulary appears in 23.7%, equally across both extraction methods.

Doctrinal vocabulary is inherent to lawyer-voice summaries: describing a slip-and-fall inevitably invokes duty and breach. Crucially, such vocabulary identifies the *topic*, not the *proposition*. The hard-negative evaluation of Appendix I shows that topical signal alone retrieves poorly, with fact-only MRR falling 65% when all negatives share the query's domain, so residual doctrinal vocabulary in the fact summaries cannot account for the gains attributed to the principle.

M Expert Annotation Template

The three legal experts worked from a common two-part instrument, reproduced here. Each annotator independently evaluated the same 15 randomly selected judgments, with an estimated 30–45 minutes per judgment.

Part A: factual field verification. Annotators verified the following fields directly against the source document: *Year*, *Court Type*, *Case Number*, *URL*, *Full Reference*, *Case Name*, *Current Court Level*, *Cited Case*, *Paragraph*, *Court Level* (of the cited case), and *Precedential Weight*. Each was graded *Correct* (matches source exactly), *Incorrect* (wrong or mismatched), or *Not Verifiable* (cannot be verified from the source document, with the reason recorded in a comments field).

Part B: AI-generated content evaluation. Annotators rated four generated fields, *Extract of Facts*, *Key Principles Illustrated*, *Issue*, and *Issue Group*, on the 1–5 scale below.

Field-specific criteria. Each field carried three written criteria. *Extract of Facts*: does it accurately capture the core factual scenario; are key parties, events and outcomes included; is it free from hallucinations or invented details? *Key Principles Illustrated*: are the legal principles accurately extracted from the cited paragraph; do they represent actual holdings or dicta from the case; are they relevant to the citing context? *Issue*: does it correctly identify the legal question addressed; is it specific enough to be meaningful; does it match the context of the citation? *Issue Group*: is the categorisation legally

Score	Rating	Description
5	Excellent	Completely accurate, comprehensive, highly relevant, well-articulated
4	Good	Mostly accurate with minor omissions or slight lack of precision
3	Acceptable	Generally correct but missing important details or has some inaccuracies
2	Poor	Significant inaccuracies or major omissions that affect understanding
1	Unacceptable	Fundamentally wrong, misleading, or completely off-topic

Table 13: Rating scale used for the AI-generated fields in Part B of the annotation template.

coherent; does it appropriately group related legal concepts; is it consistent with legal taxonomy?

Consistency instructions. Annotators were instructed to read the full cited paragraph before evaluating *Key Principles*, to check their understanding against the judgment's holding or ratio, to apply consistent standards across all 15 judgments, to prefer the middle score (3) over extremes when in doubt, to document their reasoning for any score of 1 or 5 in the comments field, and to flag any critical error that would seriously mislead a legal researcher.