Studies in Law and Justice
ISSN 2958-0382
www.paradigmpress.org/slj
Volume 5 Number 3 September 2026

ASEANLex: A Traceable Retrieval and Competency-Matching Prototype for ASEAN-Oriented Legal Education

Yunmei Li[1] & Liping Chen[1]

[1] School of Law, Guangxi Normal University, Guilin, China
Correspondence: Liping Chen, School of Law, Guangxi Normal University, Guilin, China.

doi:10.63593/SLJ.2026.09.07

Abstract

Legal-education platforms that combine document retrieval and competency-based career guidance must distinguish topically similar teaching material from reliable legal authority and differentiate high suitability scores from mandatory eligibility. This paper presents a source-code-grounded engineering diagnostic of ASEANLex, a TypeScript/Express prototype comprising a local lexical index, a coverage-based evidence gate, a five-dimension competency rubric and an explanatory demonstration-job matcher. The development collection contains 196 learning records. Five retrieval configurations were compared on 40 corpus-informed reference prompts and 22 researcher-authored unsupported probes. Ungated BM25 placed a provisional reference within the first three results for 32 of 40 prompts, against 25 of 40 for token overlap. A gate at 0.34 reduced unsupported-probe acceptance from 18 to seven of 22 while retaining support for 39 of 40 reference prompts; a 0.60 gate reduced unsupported acceptance to four but retained only 21 reference prompts. Separately, 112 constructed learner–job scores and a 45-case qualification overlay exposed five soft-pass cases without imposed mandatory-language evidence. These are reproducible development diagnostics, not expert-judged accuracy, improved student learning, or observed recruitment outcomes. We specify a blinded annotation and provenance-validation pathway for subsequent evaluation.

Keywords: legal education, information retrieval, BM25, evidence abstention, competency matching, source provenance, explainable recommendation

1. Introduction

Cross-border legal education must help learners discover jurisdiction-specific teaching resources, demonstrate relevant competencies and understand professional role requirements. Integrating these tasks in a single platform creates two distinct risks. First, a glossary entry or teaching scenario may be lexically relevant without being an authoritative statement of current law. Second, a high aggregate job-matching score can conceal a missing mandatory qualification. In both cases, a plausible software output may be mistaken for a validated decision. Lexical ranking and retrieval-augmented generation provide ways to organize material and trace possible evidence (S. Robertson& H. Zaragoza, 2009; C. D. Manning, P. Raghavan & H. Sch¨utze, 2008; P. Lewis et al., 2020). Yet a

document rank is not a legal validity certificate, and a retrieval citation is not an independent authentication of its source. Similarly, feedback on learner capabilities can aid formative reflection while remaining different from an empirically validated prediction of employment outcomes (E. Kasneci et al., 2023). We examine ASEANLex as an implemented, inspectable prototype rather than claiming a new retrieval algorithm or a tested learning intervention. The engineering contributions are: (i) a controlled comparison of ranking, evidence acceptance and collection restrictions; (ii) a diagnostic of soft qualification treatment in weighted matching; and (iii) an audit design that separates software decisions from independent human judgments. Research questions. RQ1 asks how lexical ranking, evidence thresholds and application-side verification filters affect provisional-reference placement and support assertions. RQ2 asks whether deterministic constructed-profile matching can remain stable while permitting a missing mandatory language. RQ3 asks what blinded assessment and provenance controls are necessary before such development results could become defensible quality measurements.

2. Related Work

2.1 Retrieval and Evidence Acceptance

The BM25 family ranks lexical matches using term frequency, term rarity and document-length normalization (S. Robertson& H. Zaragoza, 2009; C. D. Manning, P. Raghavan & H. Sch¨utze, 2008). Ranking quality and answer support are not interchangeable. A system may return its top-ranked item even when its collection cannot support a particular query. Retrieval augmented approaches motivate tracing outputs to passages (P. Lewis et al., 2020), but source authority, currency and jurisdictional applicability require independent assessment. In a legal learning system, a case exercise, teaching summary and authenticated instrument should therefore be presented as different kinds of material.

2.2 Capability Feedback and Eligibility

A competency rubric can summarize learning evidence across dimensions, and weighted similarity can order demonstration opportunities. A job requirement expressed as mandatory, however, is a logical constraint rather than another preference to average away. The prototype's soft scoring mechanism is consequently evaluated as a diagnostic ranking rule, not as certification of employability. Generative feedback in education also requires assessment of reliability and learner outcomes before benefits can be inferred (E. Kasneci et al., 2023).

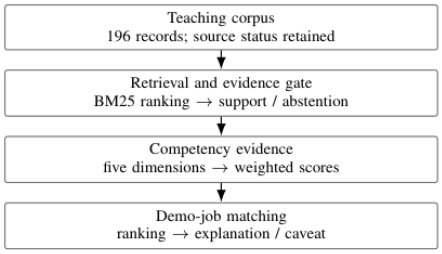


Figure 1. Code-grounded functional stages. Arrows indicate analysis flow, not independently validated legal authority or educational effectiveness.

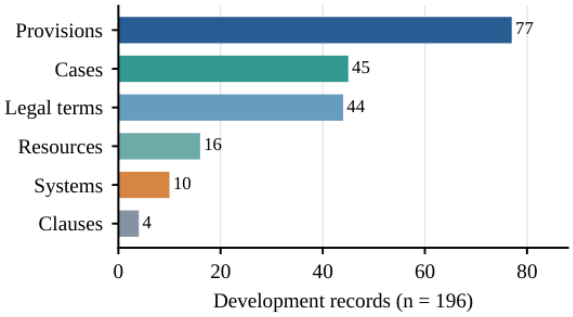


Figure 2. Composition of the 196-record development collection. Counts are database descriptors, not authenticated legal-source counts.

3. System and Evaluation Design

3.1 Implemented Pipeline and Collection

The inspected TypeScript/Express snapshot includes a local lexical index, deterministic evidence acceptance, a five-dimension rubric and a demonstration-job scoring module. Figure 1 presents analysis stages, not a legal decision chain. The development corpus comprises 77 provision records, 45 teaching cases, 44 legal terms, 16 resources, ten jurisdiction summaries and four clauses, for 196 records in total (Figure 2). Only seven items carried the application's verified flag; that internal metadata was not independently authenticated against official legal sources.

3.2 Retrieval Conditions and Acceptance Rule

Each indexed item retains an origin table and material type; the query routine returns a ranked candidate list and a support decision. Its

weighted BM25-style implementation can be understood in the family of scores

$$\text{score}(d,q)=\sum_{t\in q}\text{IDF}(t)\frac{f(t,d)(k_1+1)}{f(t,d)+k_1\left(1-b+b\frac{|d|}{\text{avgdl}}\right)} \quad (1)$$

Equation (1) is conceptual; experiments execute the compiled project code, not an independently reimplemented textbook algorithm. The subsequent lexical-coverage/evidence gate determines whether selected material is supported or whether the system should abstain. Five conditions were compared: token overlap, ungated BM25, BM25 with thresholds 0.34 and 0.60, and BM25 restricted to the seven application-verified records. The gate changes support decisions without changing the underlying BM25 order. The verified-only restriction instead changes the eligible collection.

3.3 Fixtures, Diagnostic Measures and Audit Trail

The fixed development test contains 40 queries designed with provisional reference identifiers visible to fixture authors, plus 22 researcher-written probes intended to be unsupported. Hence Anchor@1 and Anchor@3 measure placement of a designer-selected reference, not expert-judged precision. Reference accepted and unsupported accepted count implemented software support assertions on different query sets; they are not calibrated true-positive or false-positive rates. All five configurations were run on the same 62 queries, yielding 310 method–query output records with candidate IDs and support decisions. A later review pool contains 431 unique query–document pairs. Eight synthetic learner profiles paired with 14 demonstration jobs yield 112 deterministic scores. A separate 45-case qualification overlay imposes hypothetical mandatory-language constraints, deliberately including adversarial missing-language scenarios. Neither real students nor employers participated. The experimental hard-eligibility filter is a counterfactual post-processing comparison, not deployed software.

3.4 Competency and Job-Matching Logic

The matcher computes a weighted aggregate across evidenced dimensions,

$$S_j=\frac{\sum_{k\in K_j} w_{jk}s_{jk}}{\sum_{k\in K_j} w_{jk}}$$
$$M_j=0.8S_j+0.2L_j \quad (2)$$

where sjk is a constructed competence score, wjk the configured job weight and Lj the language-coverage component when configured. Because absent language evidence can reduce Mj without making it infeasible, a soft display threshold may still admit a profile that violates an imposed hard requirement. Equation (2) describes the inspected software logic; it is not an employer decision model.

4. Diagnostic Results

4.1 Reference Placement and Selective Support

Table 1 reports the counts without conflating ranking and acceptance. BM25 placed the provisional reference among the first three for 32/40 queries, versus 25/40 for token overlap. Anchor@1 was 20/40 for the ordinary BM25 conditions, versus 9/40 for token overlap. Tightening a gate did not alter the ranked list (Figure 3); it only changed whether the system asserted usable support. At threshold 0.34, unsupported-probe acceptance fell by 11 cases relative to ungated BM25 (18 to seven) while reference acceptance fell by one (40 to 39). At 0.60, unsupported acceptance fell to four, but only 21 reference prompts received support. Figure 4 makes the trade-off visible; neither series is independently adjudicated ground truth. Restricting the collection to seven internally verified records yielded only seven top-three anchor placements and one supported probe. This result concerns loss of candidate coverage, not proof that verified legal authorities perform worse.

Table 1. Code-Grounded Development Diagnostics; Not Expert Accuracy

Condition	Anchor@3(/40)	Unsupported accepted(/22)	Reference accepted(/40)
Token overlap	25	18	40
BM25, ungated	32	18	40
BM25, gate 0.34	32	7	39

BM25, gate 0.60	32	4	21
BM25, verified only	7	1	20

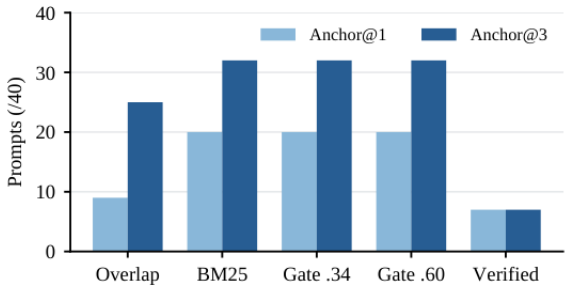


Figure 3. Provisional anchor placement, comparing rank one and top three on 40 corpus-informed prompts.

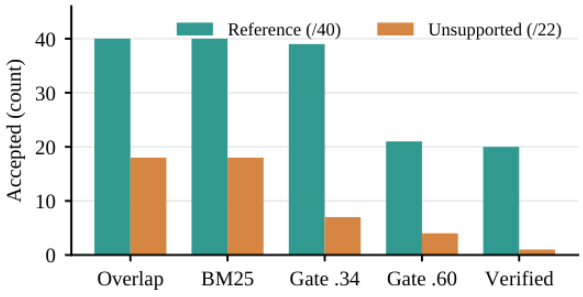


Figure 4. Support assertions on reference prompts (n = 40) and researcher- authored unsupported probes (n = 22). Denominators differ.

4.2 Descriptive Retrieval Timing

Table 2 reports local median times for four BM25 configurations; the token-overlap implementation was not timed. The observed values span 3.239–4.270 ms (Figure 6). These small implementation diagnostics were not collected as a controlled hardware-normalized benchmark; a relative speed claim would overstate their significance.

4.3 Constructed Matching and Qualification Stress Test

The 112 demonstration profile–job scores reproduced deterministically, ranging from 30 to 38.85 in the selected fixtures. Under 16 specified weight perturbations the top-ranked demonstration job stayed unchanged (16/16), with approximately 95.8% mean top-three overlap. Numerical ranking stability does not imply career-guidance validity. The separate 45-case overlay produced 16 scores above a soft display threshold of 30. Five of those 16 lacked evidence for a hypothetically mandatory language. A counterfactual hard post-filter would retain the other 11 (Table 3, Figure 5); it was not implemented in the deployed matching module. The five adverse cases were intentionally constructed rather than sampled from a student population.

Table 2. Additional retrieval descriptors

Condition	Anchor@1(/40)	Median(ms)
Token overlap	9	Not timed
BM25, ungated	20	4.031
BM25, gate 0.34	20	4.270
BM25, gate 0.60	20	3.995
BM25, verified only	7	3.239

5. Discussion And Validation Plan

5.1 Separating Ranking, Support and Authority

RQ1 shows why retrieval placement, software acceptance and source trustworthiness should be separate evaluation axes. A gate that leaves ranking invariant may still reduce unsupported support assertions. However, stricter gating also withholds assistance on corpus-informed prompts. The appropriate operating point depends on the intended educational use and the consequences of unsupported legal explanations. The seven-record verified-only restriction further illustrates that the database's status flag cannot substitute for an external legal provenance audit.

5.2 Suitability Is Not Eligibility

RQ2 is answered at the level of constructed software tests: deterministic score order may appear stable even while hypo-thetical mandatory-language violations pass a soft

threshold. A future product should expose both a suitability score and a separately evaluated eligibility state, grounded in externally checked job requirements. No conclusion about actual hiring decisions, fairness or recommendation usefulness follows from these fixtures.

5.3 Blinded Review and Failure-Closed Evaluation

For RQ3, we specify a two-reviewer protocol. Independent qualified reviewers should label all 62 queries for collection-level answerability; all 431 pooled query–document candidates for graded topical relevance and independently checked legal-source reliability; and all 112 constructed profile–job pairs for suitability and explicit eligibility. Each reviewer receives packets without model predictions or the other reviewer's labels. Disagreements are retained before adjudication so that inter-rater reliability is visible. A scoring script should reject missing or invalid labels rather than silently report partial quality metrics. With adjudicated labels, graded measures such as NDCG@5 (K. J¨arvelin & J. Kek¨al¨ainen, 2002) and verified supporting-citation proportions can be computed, while answerability errors and qualification violations should be counted separately. Since the existing 62 prompts were developed with the collection visible, even perfect post- hoc annotation would not eliminate development-set selection bias. A further held-out query collection must be authored independently, and legal authority must be authenticated for jurisdiction and revision date.

Table 3. Hypothetical language-eligibility overlay (n=45)

Condition	Cases
Meet soft threshold of 30	16
Above threshold but missing language	5
Retained by hypothetical hard filter	11
Below soft threshold	29

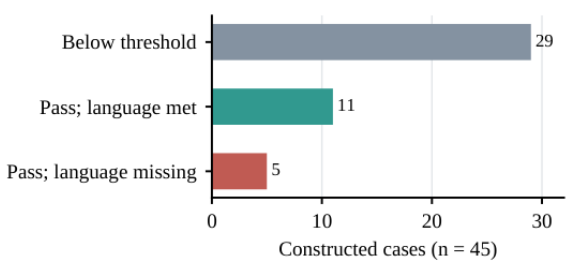


Figure 5. Qualification-overlay partition. Passing a soft score is not sufficient evidence of a mandatory language.

5.4 Threats to Validity

The strongest limitations are corpus-informed reference selection, researcher-authored unsupported probes, seven internally flagged rather than independently authenticated sources, constructed learner profiles and demonstration jobs, and the absence of actual educational or recruitment outcomes. The local timing archive also lacks sufficient controlled-environment documentation for performance generalization. There is no completed expert review of relevance or source authority, no learner comprehension test, and no field validation. Consequently, statements about real-world legal-answer accuracy, hallucination frequency, improved learning or employer acceptance are not justified.

5.5 Reproducibility Scope and Deployment Implications

The experimental record supports a comparatively narrow form of reproducibility. The fixed prompt list, five method configurations, top-five identifiers, chosen evidence fields, support flags, and constructed profile–job scores preserve what the program returned at development time. An independent replication should record the exact repository revision, executable dependency versions, normalized corpus snapshot and environment metadata, then rerun the fixtures without changing the query wording or the eligible-document filters. This would test whether the numerical outputs are stable under a frozen software environment. It would not, by itself, test whether any retrieved legal material was authoritative or any recommended job was appropriate.

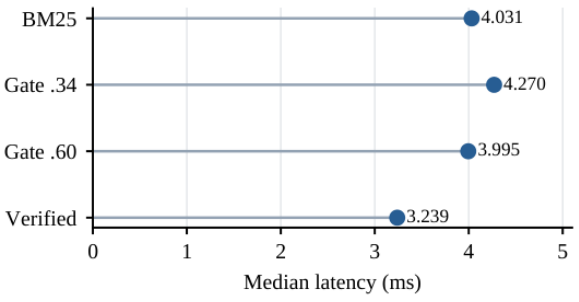


Figure 6. Descriptive median retrieval latency from the existing diagnostic runs; no hardware-

normalized inference is intended.

We recommend that future evaluations report three orthogonal dimensions. The first is retrieval relevance: given a held-out query, how well do candidate materials address the expressed information need? The second is support reliability: where the tool claims it can answer, is the selected passage sufficient and is its provenance independently authenticated? The third is educational and eligibility usefulness: can learners interpret the explanation and does a separately encoded requirement check reject mandatory-qualification failures? A single aggregate percentage would obscure these different failure modes and their different denominators.

A learner-facing deployment should also show document type, jurisdiction, original source and revision date, rather than displaying a generic citation identifier alone. The verification flag should be defined operationally, with a clear distinction between an internally curated teaching object and a primary legal instrument checked against an official source. For career guidance, the interface should distinguish "suggested fit" from "eligibility confirmed" and use the latter only after every mandatory requirement is explicitly evaluated. Where evidence is missing, a transparent unresolved state is preferable to silently treating missing information as a small numeric penalty.

Finally, any future participant study would need informed consent, a controlled comparison of feedback conditions and pre-specified education outcomes. It should also define how misleading or outdated legal explanations are detected and corrected. These requirements are beyond the evidence of the present project snapshot, but making them explicit prevents an implementation diagnostic from being interpreted as a validated classroom intervention.

6. Conclusion

ASEANLex illustrates two engineering tensions in a legal-training prototype: evidence gates can reduce support on constructed unsupported probes while sacrificing coverage, and soft competency matching can admit deliberately underqualified cases. The contribution is an auditable development diagnostic coupled with an independent validation design, not a demonstrated improvement in learning or hiring. Future work should prioritize authenticated jurisdiction-specific sources, separately enforced mandatory qualifications, blinded judgments and genuinely held-out learner tasks.

References

C. D. Manning, P. Raghavan, and H. Sch¨utze. (2008). *Introduction to Information Retrieval*. Cambridge Univ. Press.

E. Kasneci et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Art. 102274.

K. J¨arvelin and J. Kek¨al¨ainen. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), pp. 422–446.

P. Lewis et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474.

S. Robertson and H. Zaragoza. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information RetrieDval, 3*(4), pp. 333–389.