Unit 2 Independent Project
TURDIYEV ABDURAXMON BAHROM O‘G‘LI

UNIT 2: Unit 2 Independent Project

Lecturer: Gulomov Abdulaziz

Group ID: 22-307

Student ID: 220012

Submission date: 14.06.2026

BTEC Learner Assessment Submission and Declaration

When submitting evidence for assessment, each learner must sign a declaration confirming that the work is their own.

Learner (Student) ID:	220012
Assessor Name:	Gulomov Abdulaziz
BTEC Programme Title:	Pearson BTEC Level 6 Diploma in Digital Technologies (SRF)
Unit or Component Number and Title:	**Unit 2:** Unit 2 Independent Project
Assignment Title:	Legal Recommendation System for Retrieving Similar Court Cases from Judical Databases
Date Assignment Submitted:	14.06.2026

Declaration of Originality

I hereby declare that this Independent Project, submitted in partial fulfilment of the requirements for the Pearson BTEC Level 6 Diploma in Digital Technologies and the Bachelor's Degree in Business Information Technology at PDP University, is the result of my own original work.

I confirm that:

- All sources of information, data and ideas drawn from other authors have been fully acknowledged through accurate in-text citations and a complete reference list using the Harvard (author–date) referencing system.
- This work has not been previously submitted, in whole or in part, for any other academic award at this or any other institution.
- All research activities involving human participants have been conducted in compliance with the institutional ethics procedures of PDP University and applicable data-protection regulations.
- I understand that any breach of academic integrity, including plagiarism, fabrication of data or unauthorised collaboration, may result in the withdrawal of this submission and disciplinary action.

Signature:  ***Date:*** *14.06.2026*

Acknowledgements

I would like to express my sincere gratitude to my supervisor Abdulaziz Gulomov for their continuous guidance, valuable feedback and encouragement throughout the duration of this project. I am also grateful to the faculty of the Business Information Technology department at PDP University for providing me with the academic foundation and resources that made this work possible...

Finally, I extend my heartfelt thanks to my family and peers, whose support has been instrumental during every stage of my Bachelor's studies.

Abstract

This project investigates how far modern retrieval methods improve the search for relevant judicial precedent. Legal research is a hard search problem, because relevance turns on a shared legal issue, not on shared words, so keyword and Boolean tools return documents that only match terms. The aim was to design and critically evaluate a relevance-ranked retrieval system that recommends similar precedents from a plain-language description of a legal situation, and to assess how far its evaluation can be trusted.

A design-science approach was adopted. A hybrid pipeline was built over a fixed corpus of 120 United States Supreme Court opinions: semantic embeddings and a lexical channel were fused, then re-ordered by a pointwise large-language-model reranker that also assigns an absolute relevance score and can declare "no strong matches". The system was tested with a reproducible evaluation harness of 56 graded queries and 1,276 relevance judgements, scored by a large-language-model judge whose agreement was measured (kappa = 0.927).

Ranking quality, measured with nDCG@10, rose from 0.385 for keyword search to 0.817 for semantic retrieval, and to 0.902 with the reranker, a statistically significant gain. Hybrid fusion did not beat semantic retrieval alone, which corrected an initial assumption. The honest no-match signal separated in-domain from out-of-domain queries perfectly.

The project concludes that semantic retrieval and a language-model reranker measurably improve issue-level legal relevance, and that an honest relevance signal is feasible. Because the ground truth came from an automated judge rather than lawyers, the agreement statistic measures reproducibility, not validity; practitioner validation is named as future work.

Keywords: *legal information retrieval; semantic search; reranking; relevance evaluation; large language models; design science.*

Table of Contents

Declaration of Originality **2**
Acknowledgements **3**
Abstract **3**
Table of Contents **4**
List of Figures 6
List of Tables 7
List of Abbreviations 8
Chapter 1: Introduction 8
1.1 Background and Context 8
1.2 Problem Statement 9
1.3 Project Aim 9
1.4 Project Objectives 9
1.5 Research Questions and/or Hypotheses 10
1.6 Significance of the Project 10
1.7 Scope and Limitations 10
1.8 Structure of the Report 11
Chapter 2: Literature Review 12
2.1 Introduction to the Literature Review 12
2.2 Theoretical Framework 12
2.3 Thematic Review 12
2.3.1 Theme 1: Legal retrieval and the limits of lexical search 12
2.3.2 Theme 2: Dense and semantic retrieval 13
2.3.3 Theme 3: Hybrid retrieval, reranking and evaluation 13
2.4 Industry / Practice Review 14
2.5 Identification of the Gap 14
2.6 Conceptual Framework 15
2.7 Summary of the Literature Review 16
Chapter 3: Project Planning and Methodology 17
3.1 Research Philosophy and Approach 17
3.2 Methodological Choice 17
3.3 Project Management Methodology 17
3.4 Project Plan 18
3.4.1 Work Breakdown Structure (WBS) 18
3.4.2 Gantt Chart and Timeline 18
3.4.3 Milestones and Critical Path 19
3.5 Resource Planning 20
3.6 Risk Register 20

3.7 Ethical Considerations..... 22
3.8 Project Tracking and Documentation..... 22
3.9 Implementation of the Plan..... 23
3.10 Critical Assessment of Project Management..... 24
Chapter 4: Data Collection and Analysis..... 25
4.1 Data Collection Methods..... 25
4.1.1 Primary Data..... 25
4.1.2 Secondary Data..... 26
4.1.3 Sampling Strategy..... 26
4.2 Analytical Techniques..... 26
4.2.1 Quantitative Analysis..... 26
4.2.2 Qualitative Analysis..... 27
4.3 Findings..... 27
4.3.1 Demographic / Sample Profile..... 27
4.3.2 Findings: Theme / Hypothesis 1..... 28
4.3.3 Findings: Theme / Hypothesis 2..... 29
4.3.4 Findings: Theme / Hypothesis 3..... 30
4.4 Comparison of Patterns and Trends..... 30
Chapter 5: Discussion..... 31
5.1 Interpretation of Findings..... 31
5.2 Comparison with the Literature..... 31
5.3 Validity..... 32
5.4 Reliability..... 33
5.5 Project Effectiveness Evaluation..... 33
5.6 Limitations..... 34
5.7 Implications for Practice and Future Research..... 34
Chapter 6: Conclusion and Recommendations..... 35
6.1 Summary of the Project..... 35
6.2 Achievement of Objectives..... 35
6.3 Contribution to Knowledge and Practice..... 36
6.4 Recommendations..... 36
6.4.1 For Practitioners / Industry..... 36
6.4.2 For Future Researchers..... 36
6.5 Closing Remarks..... 37
Chapter 7: Reflective Evaluation of Personal and Professional Development..... 37
7.1 Choice of Reflective Model..... 37
7.2 Reflection Using Gibbs' Cycle..... 37
7.2.1 Description: What happened?..... 37
7.2.2 Feelings: What were you thinking and feeling?..... 37
7.2.3 Evaluation: What was good and bad?..... 38
7.2.4 Analysis: Why did it happen this way?..... 38

7.2.5 Conclusion: What did you learn?.....38
7.2.6 Action Plan: What will you do differently?.....38
7.3 SWOT Analysis of Personal Development.....39
7.4 Transferable Skills Gained.....39
7.5 Future Development Plan.....40
Reference List.....41
Appendices.....44
Appendix A - Project Proposal (approved version).....44
Appendix B - Full Gantt Chart.....45
Appendix C - Risk Register (full version).....45
Appendix D - Ethics Approval and Consent Forms.....45
Participant Information and Consent Form (designed instrument - not deployed).....45
Appendix E - Data Collection Instruments.....47
Section 1 - About you.....47
Section 2 - Relevance judgements (blind comparison).....47
Section 3 - The "no strong matches" signal.....49
Section 4 - Closing.....49
Appendix F - Raw / Anonymised Data Extracts.....49
Appendix G - Project Logbook / Reflective Journal Extracts.....50
Appendix H - Supervisor Meeting Records.....51
Appendix I - Assessment Criteria Mapping (for self-check).....52

List of Figures

- *Figure 2.1: Conceptual framework of the court-cases retrieval pipeline, framed by design science and legal-relevance theory* (§2.6)
- *Figure 3.1: Work breakdown structure: the six work packages of the project* (§3.4.1)
- *Figure 3.2: Project timeline (Gantt): May ideation, eval-gated June build bursts, and write-up to submission on 14 June 2026* (§3.4.2)
- *Figure 3.3: Project tracking artefacts: git commits, ADRs, design specs, the mistakes log (M001–M023) and eval gates* (§3.8)
- *Figure 4.1: S1 baseline — retrieval metrics by method on the 120-case corpus, with 95% bootstrap CIs on nDCG@10* (§4.3.2)
- *Figure 4.2: AP-R reranker OFF versus ON — overall nDCG@10 and MRR@10, and per-bucket nDCG@10* (§4.3.3)
- *Figure 4.3: Out-of-domain honesty — in-domain versus out-of-domain score separation (AUC = 1.0, TAU = 5)* (§4.3.4)

List of Tables

- *Table 3.1: Milestones and critical path* (§3.4.3)
- *Table 3.2: Resource plan* (§3.5)
- *Table 3.3: Risk register* (§3.6)
- *Table 4.1: Corpus and query-bucket profile* (§4.3.1)
- *Table 5.1: Project effectiveness evaluation by objective (RAG)* (§5.5)
- *Table 7.1: SWOT analysis of personal development* (§7.3)
- *Table 7.2: Transferable skills gained (self-assessed 1–5, with project evidence)* (§7.4)
- *Table 7.3: Future development plan* (§7.5)

List of Abbreviations

Abbreviation	Full Form
ADR	Architecture Decision Record
AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under the (ROC) Curve
BM25	Best Match 25 (probabilistic lexical ranking function)
CI	Confidence Interval
DSR	Design Science Research
FTS	Full-Text Search
GDPR	General Data Protection Regulation
HNSW	Hierarchical Navigable Small World (vector index)
IR	Information Retrieval
LLM	Large Language Model
MRR	Mean Reciprocal Rank
nDCG	Normalised Discounted Cumulative Gain
OOD	Out Of Domain
RAG	Red, Amber, Green (status rating)
RRF	Reciprocal Rank Fusion
RQ	Research Question
SCOTUS	Supreme Court of the United States
SWOT	Strengths, Weaknesses, Opportunities, Threats
TREC	Text REtrieval Conference (pooling methodology)
WBS	Work Breakdown Structure

Chapter 1: Introduction

1.1 Background and Context

Legal work is a search problem: a lawyer must find the few past decisions that govern a case, not the many that share only its words. Legal information retrieval (IR) is hard because relevance turns on a shared legal *issue*, not shared vocabulary (Van Opijnen and Santos, 2017), which keyword search optimises against. Semantic, or "dense", retrieval instead matches a plain-language description by *meaning*, and large language models (LLMs) have made these representations much stronger: in a 2024 survey, 79% of law-firm respondents expected AI to have a high or transformational effect on their work (Thomson Reuters, 2024).

Legal services are also digitising beyond the Anglo-American world where these tools matured: Uzbekistan has a large online population (DataReportal, 2024) and is moving its legal information online. The *court-cases* feature sits at this intersection: a user describes a situation in plain words and the system returns ranked precedents from a corpus of US Supreme Court (SCOTUS) decisions, in English, Russian and Uzbek.

1.2 Problem Statement

This is a complex problem in the Level-6 sense: multi-variable, uncertain and real-world. There is no single correct ranking for a natural-language query; the right answer depends on how a reader reads relevance. At its centre is *vocabulary mismatch*: someone describing a situation in plain words ("my employer fired me after I complained about safety") rarely uses the terms in the controlling opinion (retaliation, protected activity, adverse employment action). Lexical, Boolean and citation search therefore return documents that share words, not a legal issue, and that gap is where lawyer time is lost (Blair and Maron, 1985).

Two problems concern the *honesty* of what the system tells its user. An early version of court-cases fused two ranked lists with Reciprocal Rank Fusion (RRF), then showed the raw fused score (about 0.016 to 0.033) as a "relevance %" with no calibrated meaning. Worse, a ranked-list system always returns its top results, however poor: faced with an out-of-domain query (a road-traffic accident posed to a constitutional-law corpus), it shows irrelevant precedent rather than admit nothing fits. The gap is therefore twofold: a retrieval-quality gap (does meaning-based ranking beat keyword search?) and a trustworthiness gap (can it score relevance on a readable scale and say "no strong matches" when true?).

1.3 Project Aim

Aim: *To design and critically evaluate a relevance-ranked retrieval system that recommends similar judicial precedents from plain-language descriptions of a legal situation, and to assess its evaluation's validity against expert practice.*

1.4 Project Objectives

1. **O1 (review):** Critically review the academic and industry literature on legal IR, dense/semantic search, hybrid ranking and reranking to locate the gap this project addresses.
2. **O2 (design/implement):** Design and implement a hybrid (semantic plus lexical) retrieval pipeline over a 120-case SCOTUS corpus, with a pointwise LLM reranker and an honest "no strong matches" signal.
3. **O3 (evaluate):** Build a reproducible evaluation harness (TREC-style pooling, graded 0–3, LLM-as-judge with inter-judge agreement, bootstrap confidence intervals) and measure retrieval quality across query types.
4. **O4 (validate + recommend):** Critically evaluate the validity and reliability of the evaluation evidence, characterise how far it can be trusted against expert practitioner judgement, drawn **from the literature** (named as future work; no survey was conducted), and produce evidence-based recommendations.

1.5 Research Questions and/or Hypotheses

The project is a hybrid capstone, so it is framed by research questions, not a hypothesis.

RQ1: To what extent does semantic and hybrid retrieval improve the relevance of recommended precedents over lexical (keyword) search for plain-language legal queries?

RQ2: Does a pointwise LLM reranker with absolute relevance scoring further improve ranking quality, and can it reliably signal when no strongly-relevant precedent exists (out-of-domain honesty)?

RQ3: How valid and reliable is an LLM-as-judge evaluation of legal relevance relative to practitioner judgement, and what does this mean for trusting the reported metrics?

1.6 Significance of the Project

The academic significance is that the project measures the two gaps from 1.2 rather than asserting them: much work applies dense retrieval and LLMs to legal text, but far less isolates how far relevance improves over a keyword baseline or treats the relevance signal's trustworthiness as a result. Its ground truth is an LLM judge, not lawyers, so RQ3 asks what an inter-judge agreement statistic can justify.

Industry significance is practical: surfacing on-issue precedent first reclaims reading time, the scarcest resource, and the cross-lingual entry point lets a Russian- or Uzbek-speaking researcher query English-language case law without translating the question. The honest "no strong matches" signal is most likely to transfer, since refusing to fake confidence on out-of-domain queries is safer for non-experts than always answering.

For the author, the significance is professional. The topic aligns with a larger legal and regulatory-technology venture for which JurisBase is the data and retrieval foundation. Building and evaluating court-cases deepened the author's grasp of how case-law material is structured, retrieved and judged for relevance, and sharpened a professional skill: using agentic-AI software

engineering critically, accepting, rejecting and verifying its output, to build and check a non-trivial system.

1.7 Scope and Limitations

In scope is the court-cases work alone, which writes only to the court_* tables: the hybrid retrieval pipeline, the AP-R pointwise LLM reranker and the honest no-match signal defined in the objectives above, with a reproducible evaluation harness over a 120-case SCOTUS corpus from CourtListener. **Out of scope** are the separate JurisBase law pipeline (BGE-M3, never touched), the production interface beyond the query path, learning-to-rank and GPU-hosted rerankers, and any AI-generated or fictitious cases.

Four limitations bound the work. A small budget and no Cohere Rerank put a gpt-4o-mini reranker in place of a commercial one. Usable Uzbek case law was unavailable, so US precedent was the pragmatic choice and a local corpus is future work. The 120-case corpus suits an overall metric but is too homogeneous for a per-query-type headline, so at-scale numbers are deferred to a roughly 1,000-case corpus. Most seriously, the relevance ground truth is an LLM judge, not human experts, so inter-judge agreement measures the labels' *reproducibility*, not their validity against practitioner judgement (mistake M009).

1.8 Structure of the Report

Chapter 2 reviews the literature and weighs the alternatives. Chapter 3 covers methodology and project management via the research onion. Chapter 4 reports results across keyword, semantic, hybrid and reranked conditions. Chapter 5 examines validity and reliability, especially the LLM-as-judge ground truth. Chapter 6 assesses each objective and recommends; Chapter 7 reflects on professional development.

Chapter 2: Literature Review

2.1 Introduction to the Literature Review

This review places *court-cases* between two fields that rarely meet: how legal documents are retrieved, and recent work on dense and hybrid neural retrieval. The search drew on Google Scholar, the ACL Anthology, arXiv, the ACM and IEEE libraries, SSRN, and *Artificial Intelligence and Law*, combining three term clusters:

- **Legal terms:** legal information retrieval, precedent retrieval, relevance in law.
- **Method terms:** dense passage retrieval, BM25, reciprocal rank fusion, cross-encoder reranking, LLM reranker.
- **Evaluation terms:** nDCG, pooling, LLM-as-judge.

A source was included if peer-reviewed or the original statement of a method used here; vendor marketing counts as evidence of practice, never effectiveness, and unverifiable claims were excluded.

2.2 Theoretical Framework

Two lenses frame this project, because it is both a build and an evaluation. The build is framed by design science: Hevner et al. (2004) describe creating and evaluating useful IT artefacts that solve real problems, and Peffers et al. (2007) turn this into a six-step process from problem identification to communication that maps onto this project. Design science then asks whether the artefact actually solves the stated problem.

Evaluation needs a second theory, because design science says an artefact must be tested but not what "good retrieval" means in law. Van Opijnen and Santos (2017) argue legal relevance has several dimensions at once: a document can be algorithmically relevant (matching query words), bibliographically relevant (procedurally or doctrinally linked), or substantively relevant (bearing on the legal issue), and these diverge. This breaks keyword search's assumption that word overlap stands in for relevance, and justifies grading against issue-level relevance. The lenses split the work: design science governs how the artefact is built and tested, legal-relevance theory what counts as success.

2.3 Thematic Review

2.3.1 Theme 1: Legal retrieval and the limits of lexical search

Blair and Maron (1985) had lawyers search a 40,000-document corpus with a strong Boolean system: they believed they were retrieving about three-quarters of the relevant material, but only about 20% surfaced, and, never seeing what they missed, they could not tell. Van Opijnen and Santos (2017) explain why this is structural: when relevance turns on a shared issue stated in doctrinal language a lay enquirer would not use, ranking by word overlap optimises the wrong target. Schafer and Maxwell (2010) offer query expansion as a partial fix that trades precision for recall and never closes the lay-to-legal gap.

These critiques agree on the problem, not the cure: failed recall and over-confidence (Blair and Maron, 1985), a language problem for NLP (Schafer and Maxwell, 2010), or a conceptual one, since no single signal captures multi-dimensional relevance (Van Opijnen and Santos, 2017). None answers the question here: whether a learned semantic representation closes the gap keyword expansion only narrows. Susskind (2023) adds that incumbent tools are overdue for change, a forecast treated as context, not evidence.

2.3.2 Theme 2: Dense and semantic retrieval

Dense retrieval answers vocabulary mismatch by encoding query and document as vectors whose closeness reflects meaning, not shared words. Reimers and Gurevych (2019) made this practical at sentence level with Sentence-BERT. Karpukhin et al. (2020) extended it to passage retrieval; their dual-encoder (DPR) retrieved relevant passages far more often than BM25. This SBERT-to-DPR line is the ancestor of the embedding channel used here.

Dense retrieval is not a settled win. Izacard et al. (2022) showed with Contriever that it can be trained without labels through contrastive learning, beating BM25 on most BEIR datasets, which matters because labelled legal judgements are scarce. But BEIR cuts the other way: Thakur et al. (2021) found supervised dense models often lose to a well-tuned BM25 under domain shift, while BM25 stays robust. So a model trained on web question-answering cannot be assumed to transfer to judicial prose, and the prudent design hedges rather than commit to dense-only retrieval. The embedding model here is OpenAI's text-embedding-3-small (OpenAI, 2024), chosen on its MTEB-leaderboard standing (Muennighoff et al., 2023). But MTEB is general-purpose with no SCOTUS-style task, so a high rank signals general quality, not legal fitness.

2.3.3 Theme 3: Hybrid retrieval, reranking and evaluation

If neither lexical nor dense retrieval wins reliably across domains, the move is to combine them. The lexical part rests on BM25, the mature term-weighting model dense methods are measured against (Robertson and Zaragoza, 2009). Cormack, Clarke and Büttcher (2009) showed that Reciprocal Rank Fusion, fusing rankers by the reciprocal of their ranks, beats more elaborate learned methods with no calibration or training. But a hybrid does not always beat its best single part: fusing a weak ranker with a strong one can drag the result down (Thakur et al., 2021), so whether fusion helps on a given corpus is an empirical question this project tests rather than assumes.

Reranking offers a second gain by re-scoring a shortlist with a heavier model. A BERT cross-encoder lifts ranking well above first-stage retrieval, at a cost that limits it to a top-k shortlist (Nogueira and Cho, 2019), and monoT5 recasts the task generatively, generalising better when data is scarce (Nogueira et al., 2020). Sun et al. (2023) then showed with RankGPT that a general LLM, prompted to rank, rivals supervised rerankers with no task-specific training; it is the direct ancestor of this project's pointwise LLM reranker. Cross-encoders and monoT5 need domain training data this project lacks, whereas an LLM reranker needs none but pays in per-query inference cost and the fragility of prompt-based scoring.

Evaluation is where the project's honesty rests. Järvelin and Kekäläinen (2002) introduced normalised discounted cumulative gain (nDCG), the graded, rank-sensitive metric reported here; because graded judgements cannot be produced at full scale, the field pools

candidates, and Buckley and Voorhees (2000) study how many queries a trustworthy comparison needs. Where more than one assessor judges, Cohen (1960) supplies the kappa statistic. The most contested move replaces human assessors with a model: Zheng et al. (2023) report a strong LLM judge agreeing with humans at about the level humans agree with each other (around 80%), attractive for a project with no panel of lawyers, but list its biases of position, verbosity and self-preference. The verdict is conditional: an automated judge is usable only if its agreement is measured and failure modes checked, not trusted by default.

2.4 Industry / Practice Review

Commercial legal research has been led for decades by Westlaw and LexisNexis, whose value lies less in raw search than in curated citation networks, headnotes and key-number taxonomies layered over Boolean retrieval. That model is now being rebuilt around generative AI: Thomson Reuters (2023) acquired Casetext, whose GPT-4-based CoCounsel assistant it treats as core, for about USD 650 million. The scale shows in its *Future of Professionals* report, where around 79% of law-firm respondents expect a high or transformational AI impact (Thomson Reuters, 2024), though the American Bar Association's (2023) *TechReport* finds adoption among smaller firms uneven. The open-data counterpoint is CourtListener (Free Law Project, 2024), which publishes more than nine million US decisions and is the corpus source here, supplying the raw material but none of the editorial relevance signal the citators sell.

Against this landscape, *court-cases* is neither a citator nor a CoCounsel rival. The commercial tools serve a curated, US-practitioner workflow and keep their ranking logic proprietary, their effectiveness asserted rather than independently measured in public. This project instead applies the published methods reviewed above to the open CourtListener corpus and tests them with a transparent, reproducible evaluation. Its contribution is not to beat a black box but to measure, in the open, what a hybrid-plus-rerank pipeline achieves on a fixed legal corpus, and how far that evidence reaches.

2.5 Identification of the Gap

Read together, the three themes expose a specific gap. The retrieval methods are mature, but the evidence sits in domains that are not law: DPR and Contriever are validated on question-answering and web benchmarks (Karpukhin et al., 2020; Izacard et al., 2022), and BEIR warns such results may not survive domain shift (Thakur et al., 2021). The legal-IR literature shows relevance in law is multi-dimensional and issue-level (Van Opijnen and Santos, 2017) but does not measure how far modern neural retrieval closes the vocabulary-mismatch gap Blair and Maron (1985) first quantified. So issue-level relevance gains from a documented hybrid-plus-rerank pipeline on a *fixed* legal corpus are under-reported. A second, sharper gap concerns the relevance *signal* itself: a list-ranking system always returns a top-k however poor the matches, yet the reviewed work almost never treats the absolute relevance score as a first-class evaluation target, rarely asking whether it is calibrated or whether the system can honestly say "no strong matches" on an out-of-domain query. The project addresses both: it measures issue-level relevance on a 120-case SCOTUS corpus using nDCG against graded judgements, and treats honest no-match behaviour and the validity of the relevance signal as outcomes to be tested, not assumed.

Reaching this position meant rejecting several plausible alternatives, each for a reason the literature supports. A *pure-keyword baseline* was rejected because Blair and Maron (1985) already establish its recall ceiling; it is kept only as the comparison floor. A *pure-dense* approach was rejected because BEIR shows dense models can lose to BM25 under domain shift (Thakur et al., 2021), and a SCOTUS corpus is exactly such a shift, which is why retrieval here is hybrid. *Benchmarking the commercial black boxes* (Westlaw, CoCounsel) was rejected for reproducibility: their ranking logic and corpora are proprietary, so any comparison would be neither repeatable nor attributable to a known method, defeating the design-science demand for transparent evaluation (Hevner et al., 2004). A *learning-to-rank or GPU cross-encoder reranker* (monoT5; Nogueira et al., 2020) was rejected on feasibility, since both need domain-labelled training data and hardware the project lacks, whereas a prompt-based LLM reranker (Sun et al., 2023) needs neither and trades those costs for per-query inference. Finally, a *purely qualitative study* of practitioner perceptions was rejected because it cannot answer RQ1 and RQ2, which are quantitative ranking questions; practitioner judgement is a future validity check (RQ3), not a claim made here. The chosen direction follows from the evidence, the feasibility limits and the two framing theories: a hybrid retriever, an LLM reranker with an honest no-match signal, and an LLM-as-judge evaluation whose own agreement is measured (Cohen, 1960; Zheng et al., 2023).

2.6 Conceptual Framework

Figure 2.1 ties the reviewed ideas into the pipeline this project evaluates. A plain-language query is encoded by the embedding model (OpenAI, 2024), the answer to Theme 1's vocabulary mismatch, then enters a hybrid stage fusing the dense channel with BM25 by reciprocal rank fusion (Cormack, Clarke and Büttcher, 2009), hedging Theme 2's domain-shift caveat (Thakur et al., 2021). A pointwise LLM reranker after Sun et al. (2023) assigns an absolute relevance score driving the honest-relevance gate, which declares no strong matches where warranted rather than forcing a misleading top-k. The precedents are judged against issue-level relevance (Van Opijnen and Santos, 2017) and measured with nDCG (Järvelin and Kekäläinen, 2002), with the framing theories of §2.2 underpinning the pipeline.

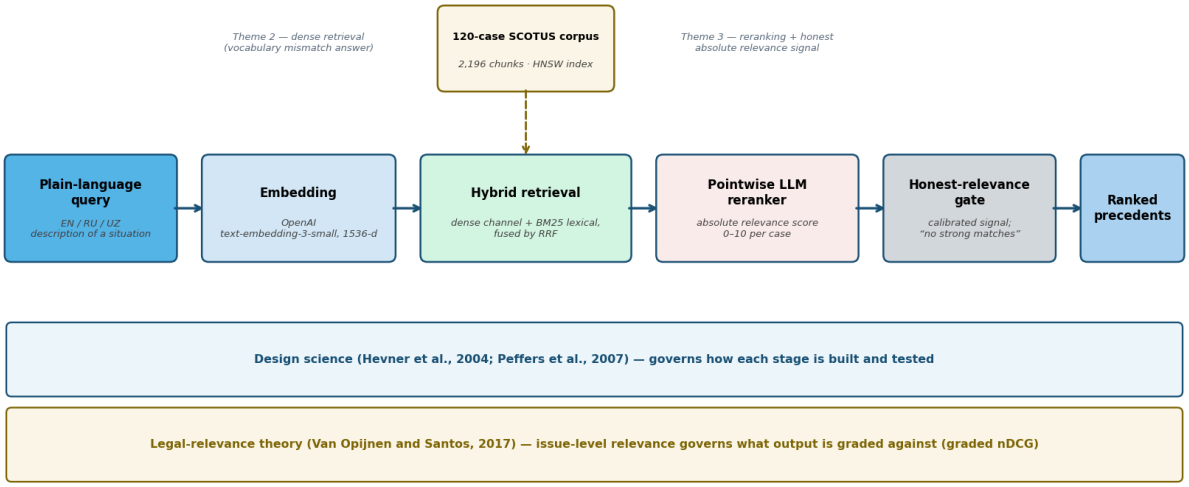


Figure 2.1: Conceptual framework of the court-cases retrieval pipeline, framed by design science and legal-relevance theory

2.7 Summary of the Literature Review

The literature establishes four commitments this project carries forward. First, keyword and Boolean search structurally under-serve issue-level legal relevance (Blair and Maron, 1985; Van Opijnen and Santos, 2017). Second, dense retrieval is the leading response to vocabulary mismatch (Karpukhin et al., 2020), but BEIR's domain-shift warning means it cannot be trusted blind on a legal corpus (Thakur et al., 2021); hence a hybrid with an LLM reranker, not dense-only. Third, two assumptions are unsafe, that a hybrid always wins and that an automated judge can be trusted by default, so each is tested empirically, the judge's agreement measured with kappa (Cohen, 1960) and quality with nDCG (Järvelin and Kekäläinen, 2002). Fourth, the gap is twofold: under-reported issue-level relevance gains on a fixed legal corpus, and the untested trustworthiness of an absolute relevance signal. These set the terms of Chapter 3.

Chapter 3: Project Planning and Methodology

3.1 Research Philosophy and Approach

This project builds and judges an artefact that ranks judicial precedents from a plain-language description of a legal situation. Neither pure positivism nor pure interpretivism fits work built to be useful then tested with numbers, so the stance is **pragmatism**: the research question decides the method, and a claim is judged by whether it works in practice (Saunders, Lewis and Thornhill, 2023). The reasoning is **abductive**, moving between a literature theory (semantic and hybrid retrieval should beat keyword search) and the system's data, sharpest at the S1 evaluation where Reciprocal Rank Fusion (RRF) did not beat semantic-only retrieval, against an assumption already in production (M007).

The strategy is **design science research** (DSR), which takes building and evaluating an artefact as its unit of inquiry (Hevner et al., 2004). Its three-cycle model fits: a relevance cycle (finding on-point precedent without reading dozens of dockets), a design cycle (the build-and-test loop over the three-table court_* schema and ranking stack), and a rigour cycle drawing on Chapter 2. The work followed the six activities of Peffers et al.'s (2007) Design Science Research Methodology (§3.4, §3.9).

3.2 Methodological Choice

A **mixed-methods** design was chosen, led by a quantitative strand and completed by a lighter qualitative one (Creswell, 2014). The quantitative core is a retrieval evaluation: 56 hand-authored queries judged against the 120-case corpus to give 1,276 graded relevance pairs, scored with standard metrics (nDCG@10, MRR@10, recall@10, precision@5) and 95% bootstrap confidence intervals. RQ1 and RQ2 are comparative and ordinal, so only a metric with a confidence interval answers "to what extent".

A quantitative result alone cannot answer RQ3: how far an LLM-as-judge can stand in for a practitioner's judgement of relevance. That is about meaning, so a modest qualitative strand remains: the legal-relevance synthesis in Chapter 2, plus informal proxy-user feedback during the build, treated as formative input, not evidence. A full thematic analysis (Braun and Clarke, 2006) of practitioner rationales was designed as the primary RQ3 validity check, but no survey ran in the project window, so that arm is future work (§3.7).

3.3 Project Management Methodology

The project was managed in an **Agile, iterative** way rather than by Waterfall: the requirement that mattered, which ranking method returns relevant results on this corpus, was unknown until measured, and the first confident answer (RRF hybrid) proved wrong.

Discipline came from the project's rule, **"eval-first, then optimise"**: trust the measuring instrument before changing anything it measures. Every ranking change had to clear an **eval-gate**, shipping only if its nDCG@10 was at least the current baseline with the confidence interval supporting the move, and rolled back otherwise.

The build ran through an **agentic-coding workflow**, itself a deliberate method choice: the system and harness were developed by directing AI coding agents (Claude Code), with the author setting each task, then reviewing, accepting, rejecting or re-running the output. This is justified like any method choice, grounded in the literature on AI-assisted software engineering (Ziegler et al., 2024), and is a strength only because the author stayed in command of the eval-gate and design decisions. Decisions were logged as the work happened, in ADRs, a mistakes log and dated commits, so the dissertation could be written in parallel.

3.4 Project Plan

3.4.1 Work Breakdown Structure (WBS)

The project breaks into six work packages, each a deliverable rather than an activity, so "done" can be defined and checked, a discipline forced by M003, where "done" had meant "committed" while retrieval was silently broken (Figure 3.1).

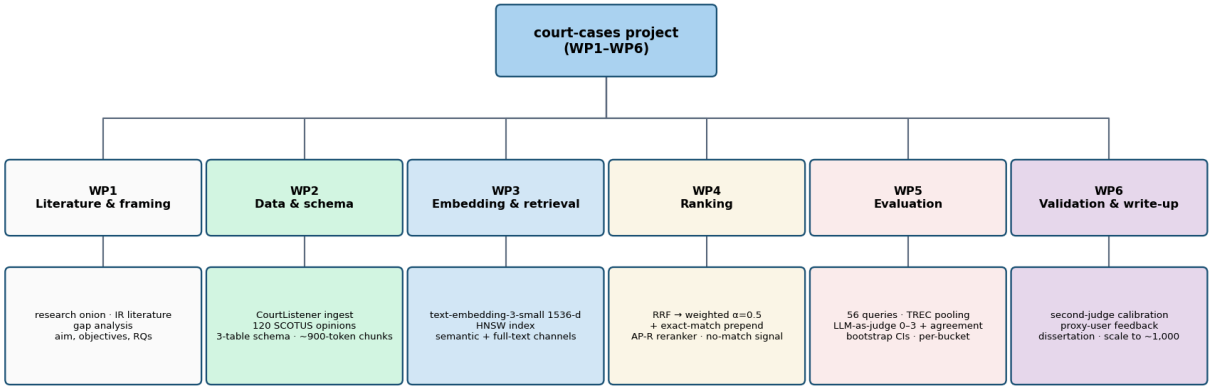


Figure 3.1: Work breakdown structure: the six work packages of the project

- **WP1 Literature and framing:** research onion, IR literature, gap analysis, aim, objectives, RQs.
- **WP2 Data and schema:** CourtListener ingest of 120 SCOTUS opinions; three-table schema; chunking (~900 tokens, ~12% overlap, tiktoken).
- **WP3 Embedding and retrieval:** OpenAI text-embedding-3-small (1536-d), HNSW index, semantic and full-text channels.
- **WP4 Ranking:** RRF, then weighted fusion (α=0.5) with an exact-match prepend, then the AP-R pointwise reranker and honest "no strong matches" signal.
- **WP5 Evaluation:** 56-query set, TREC-style pooling, graded 0–3 by an LLM-as-judge, inter-judge agreement, bootstrap CIs, per-bucket metrics.
- **WP6 Validation and write-up:** calibration against a second judge, informal proxy-user feedback, dissertation, scaling toward ~1,000 cases.

3.4.2 Gantt Chart and Timeline

The schedule began in early May 2026 with scoping, then ran as a few dated, eval-gated bursts to submission on 14 June 2026, with contingency as slack between bursts rather than a block at the end (Figure 3.2).

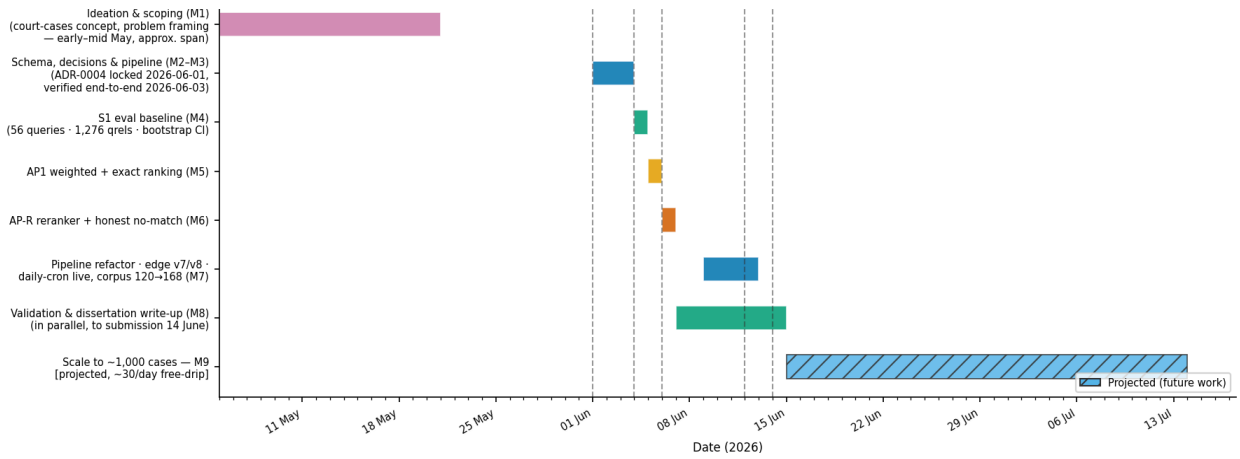


Figure 3.2: Project timeline (Gantt): May ideation, eval-gated June build bursts, and write-up to submission on 14 June 2026

3.4.3 Milestones and Critical Path

The critical path runs WP2 → WP3 → WP5: ranking quality cannot be claimed until the corpus is ingested, embedded and measurable. The evaluation instrument (milestone M4) therefore gates every later ranking milestone, each of which shipped only after passing the eval-gate (Table 3.1).

Table 3.1: Milestones and critical path

#	Milestone	Target Date	Status
M1	Scoped: court-cases concept, problem framing, initial research	Early–mid May 2026	✔
M2	Schema and Phase-1 decisions locked (ADR-0004; OpenAI 1536-d)	2026-06-01	✔
M3	Pipeline verified end-to-end (2,196 chunks; browser smoke test)	2026-06-03	✔
M4	S1 baseline established (56 queries, 1,276 qrels, bootstrap CI)	2026-06-04	✔
M5	AP1 weighted + exact-match ranking, under eval-gate	2026-06-05	✔
M6	AP-R reranker + absolute relevance + honest “no strong matches”	2026-06-06	✔
M7	Pipeline refactor (parity-gated), edge v7/v8, daily-cron live (corpus 120→168)	2026-06-09…12	✔
M8	Deployed; validation and dissertation completed	2026-06-13/14	✔
M9	Headline metrics recomputed on ~1,000-case corpus (cron live, ~30/day)	~2026-07 (projected)	⚠ future work

3.5 Resource Planning

The resource base was lean (Table 3.2), and that leanness shaped two decisions: no Cohere budget forced the gpt-4o-mini reranker, and the small OpenAI budget kept evaluation on the 120-case corpus, scaling deferred.

Table 3.2: Resource plan

Resource Type	Description	Status
Data and storage	Supabase **lexportal** (Postgres, pgvector, HNSW, Deno/TypeScript edge functions); CourtListener API v4 upstream	Available
AI services	OpenAI text-embedding-3-small (1536-d) embeddings; gpt-4o-mini reranker, topic classification, cheap-judge test; Claude Opus LLM-as-judge, Claude Sonnet reproducibility validator	Available (small budget)
Development workflow	Claude Code agentic-coding: custom agents and skills, directed and verified by the author	Available
Backend	Python inspector_x in jurisbase_backend/; standalone ingest and eval scripts (not Worker jobs, ADR-0003)	Available
Frontend	Vite + React + shadcn (jurisbase_frontend/, dev :8080/cases)	Available
Project control	git conventional commits; four ADRs, six design specs; mistakes/ log; status and handoff docs	Available
Human	Supervisor guidance (PDP University); colleagues as proxy users for informal feedback	Available
Financial	Cheap-judge run **$0.0157**; topic classification **$0.011**; query-time reranker cost tracked, not yet projected	Tracked

3.6 Risk Register

Likelihood and impact are scored 1–5; the residual RAG reflects the position *after* mitigation (Table 3.3). Several listed risks materialised, logged in mistakes/.

Table 3.3: Risk register

#	Risk Description	L (1–5)	I (1–5)	Score	Mitigation Strategy	Residual RAG
R1	120-case corpus too small/homogeneous (one court, constitutional-heavy) for a headline metric	5	4	20	120 for development/tuning only; report relative comparisons (bootstrap CIs, per-bucket); defer headline to ~1,000 cases	Amber

#	Risk Description	L (1–5)	I (1–5)	Score	Mitigation Strategy	Residual RAG
R2	LLM-as-judge is not human ground truth; absolute metrics reflect the judge, not legal relevance (M009)	4	4	16	Validate reproducibility (κ = 0.927 vs a second LLM); claim only *relative* comparisons; practitioner validation as future work	Amber
R3	OpenAI budget and rate limits constrain embedding and reranking at scale	4	3	12	Batch embeddings (128); cheaper gpt-4o-mini reranker; cost-track every run; cap reranker at K=20	Green
R4	CourtListener free-tier limit delays corpus scaling	4	3	12	Background cron with contingency buffer; cache aggressively; idempotent ingest (~30 cases/day)	Amber
R5	Breaking court↔law isolation (ADR-0003/0004): a court change corrupts live law_chunks / BGE-M3	2	5	10	Hard rule: court writes only court_*; standalone scripts, never a SupabaseWorker job; reviewed in every ADR	Green
R6	Solo build plus parallel AI sessions on one repository: lost work, wrong status, branch races (M001, M011, M020)	4	3	12	Anchor status to a source of truth; one branch per work-stream; push fast-forward-only; isolated checkout per session	Amber
R7	Config/secret drift across surfaces breaks the deployed system (M005 unset edge secret; M006 dependency drift)	4	3	12	Full-surface checklist for any provider/key change (backend env, edge secret, .env.example, docs); treat pyproject.toml/.env.example as contracts	Green

3.7 Ethical Considerations

The case corpus is public United States court data from CourtListener: published opinions with no personal-data concern, so the only duty is to respect its open-data terms and cite it. The one human input was informal verbal feedback from colleagues acting as proxy users; no personal data was recorded, they are not legal experts, and their comments are reported as formative input, never as survey results. With no human-subject data collected, no ethics-committee approval was needed. A practitioner survey was designed as the primary external validity check for RQ3 but not deployed in the project window; its instrument and consent form (Appendices E and D) were built to the standard a real study needs and are presented as designed-not-deployed, recording planning rigour without claiming data that does not exist.

A second matter is specific here: parts of the system and harness were built with AI assistance. AI was a construction tool, not an author of the research claims: it wrote pipeline and harness code, but not the 56 evaluation queries. Where an LLM acts as the relevance judge it is named as such, its reproducibility reported as $\kappa = 0.927$ (Cohen, 1960) and its limits stated openly (Zheng et al., 2023). The switch from planned manual labelling to an LLM-as-judge is disclosed as a methodological deviation (M009, §3.10), not hidden.

3.8 Project Tracking and Documentation

The tracking system was the engineering record itself, not a commercial board, anchored to artefacts rather than self-report. Four instruments carried it (Figure 3.3): **git** with conventional commits, for a dated, auditable trail; **four ADRs and six design specs**, recording each non-trivial decision with its rationale and rejected alternatives; the **eval-gated task lists** (PASS, FAIL or NEUTRAL, with rollback), turning “track progress” into “track whether each change cleared its evidence bar”; and the **mistakes/ log (M001–M023)**, a defect-and-lesson register that §3.10 draws on.

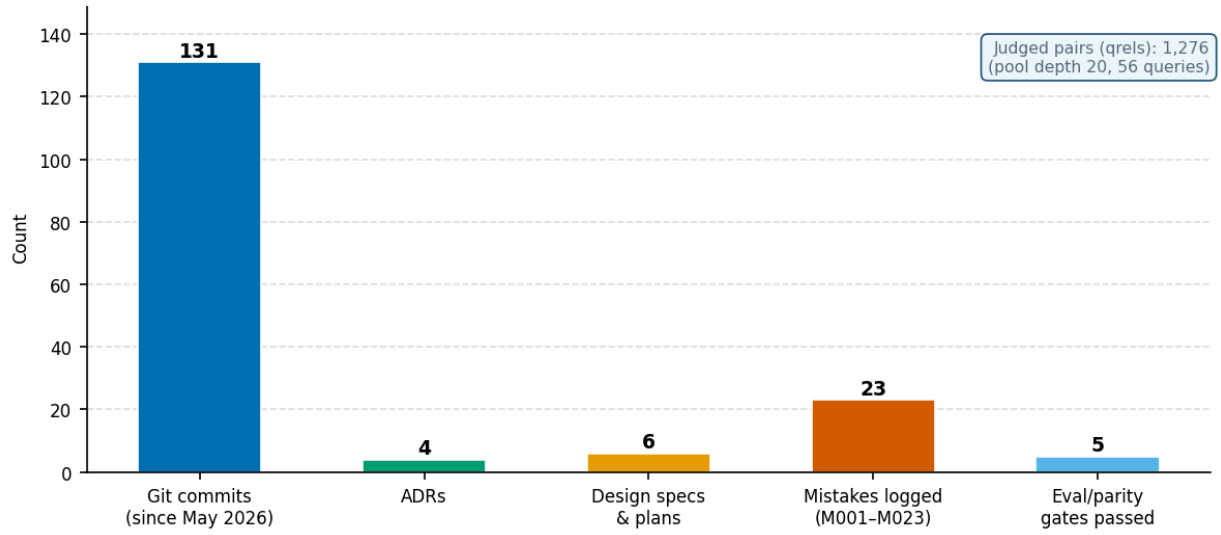


Figure 3.3: Project tracking artefacts: git commits, ADRs, design specs, the mistakes log (M001–M023) and eval gates

The record also shows the project *responding* to change, the half of M2 a static board cannot evidence. The S1 baseline disproved the production hybrid and turned the plan toward reranking (M007). A partial-ingest bug that would have polluted the full-text channel was caught by code inspection before reaching the eval corpus (M010). An edge secret found unset after the provider switch prompted a config checklist (M005). When the daily-cron added cases with missing decision dates, a wrong diagnosis and a quick guard broke it, until a live source check corrected both (M022, M023). These artefacts were how the project steered, not after-the-fact documentation.

3.9 Implementation of the Plan

The plan executed in the DSR sequence and largely held to schedule (Table 3.1). WP2–WP3 made the 120 SCOTUS opinions searchable under ADR-0004, and the pipeline was verified end-to-end once a backfill produced 2,196 chunks with 1,536-dimension embeddings and a browser smoke test returned live results. WP5 built the measuring instrument: the S1 baseline over 56 queries and 1,276 judged pairs with bootstrap CIs. Under the eval-gate, AP1 replaced RRF with weighted fusion plus an exact-match prepend (nDCG@10 0.841 against RRF's 0.784), lifting the exact bucket from 0.698 to 0.951; AP-R then added the gpt-4o-mini pointwise reranker and the honest no-match signal (nDCG@10 0.902, a significant +0.085 over semantic, no per-bucket regression, no pool leakage).

Three deviations show the gate working. First, the semantic channel was averaging chunk scores; switching to a best-chunk (MAX) score was corrected before any method comparison was trusted. Second, RRF had shipped as the assumed-best hybrid, and S1 contradicted that (M007), so the plan demoted RRF to weighted fusion and layered the reranker. Third, the daily-cron later scaled the live corpus from 120 to 168 cases, surfacing a date-mapping bug and a misdiagnosis (M022, corrected by M023); the eval stayed frozen at 120, so no reported number moved. Throughout, the AI-assisted workflow was directed, not trusted: agent output was checked against the eval-gate and tests, and rolled back when it failed them.

3.10 Critical Assessment of Project Management

The most effective decision was the eval-before-ship gate, proven by its catching the project's largest error. RRF hybrid had shipped as the production default on the textbook theory that hybrid beats semantic. The S1 evaluation disproved it on this corpus (M007), and only the gate stopped a wrong assumption becoming a permanent claim. Without it, the dissertation would have reported a hybrid system as its best result and been quietly wrong, the clearest sign the measurement-first management worked.

It is equally honest to name where management failed. Status was often inferred, not verified: the court tables were called "empty" from a missing checkpoint file when the live database held 120 full-text opinions, because no count query was run (M001); the same "done meaning committed" pattern called a feature complete while its chunk table held zero rows (M003). Configuration discipline lapsed twice: an edge secret unset after the OpenAI switch (M005), and an undeclared dependency across mismatched environments (M006). A wider habit ran through later work, trusting a plan or status field over a runtime check: a deploy plan called ready without tracing its import graph (M013); scale-ingest tooling called "verified" until a dry-run found four defects (M016); a minimal cron image reasoned from a full development

environment until it failed in a clean one (M021); and the date-mapping misdiagnosis (M022), an empty column read as “undecided” without checking the source.

Three improvements follow, each tied to a failure above. First, **status must be anchored to a source of truth, not an artefact**: a count query or smoke-test output, never the presence or absence of a file, a real hazard working solo, where no peer challenges a claim. Second, **“done” needs an operational definition per deliverable**, checked in an environment that mirrors production (the lesson M013, M016 and M021 share). Third, **any cross-surface change** (backend, edge secret, example files, docs) **needs a checklist**, because a provider switch is never a single edit (M005, M006). One lesson matters more for research than for the build: a change to the evidence-generating method, such as swapping manual labels for an LLM judge, must be recorded in the methodology as it happens, with its effect on what can be claimed (M009). Running build and write-up in parallel met the deadline but raised the risk that materialised as parallel sessions racing on one repository (M020); next time, each work-stream gets its own checkout.

Chapter 4: Data Collection and Analysis

4.1 Data Collection Methods

The evidence comes from two streams. The primary stream is generated by the system under controlled conditions: a relevance evaluation in which hand-authored queries are run against the corpus and every retrieved candidate is graded. The secondary stream is the legal corpus itself, drawn from a public archive. A third instrument, a practitioner survey, was designed as the external check but not deployed inside the project window; it is described in §4.2.2 and §3.7 as a future-validation tool held in Appendices D and E, not as data reported here.

4.1.1 Primary Data

The primary data is the relevance evaluation, which needs queries and, for each, a verdict on which documents are relevant. Fifty-six queries were hand-authored across five buckets, each stressing a different retrieval behaviour: 16 fact_pattern (a plain-language situation, the central use case), 12 paraphrase (a holding restated in lay terms), 10 short (two or three keywords), 10 exact (a case name or citation), and 8 cross_lingual (5 Russian, 3 Uzbek). Each bucket isolates a failure mode a single aggregate would hide; the cross-lingual bucket exists because the intended users sit in Uzbekistan and search an English-language corpus.

Judging every query against all 120 cases would mean 6,720 mostly-trivial verdicts, so the evaluation uses **TREC-style pooling**: for each query the top 20 results from every method are merged into one pool, the producing method hidden, and only pooled candidates graded (Buckley and Voorhees, 2000). Each (query, opinion) pair was graded 0–3, giving **1,276 judged pairs**, steeply skewed (0 = 914, 1 = 240, 2 = 63, 3 = 59) as expected where most candidates are off-topic.

The judgements are not human. The original plan, a student hand-labelling roughly 36 queries, was overridden for an **LLM-as-judge** design, recorded as mistake M009 because a change of instrument deserved an explicit decision. The primary grader is **Claude Opus**, run as eight parallel agents against a fixed rubric anchored to what a practising lawyer would want. An LLM judge is defensible, since GPT-4-class models reach roughly 80% agreement with human raters on comparable tasks (Zheng et al., 2023), but it is not self-validating. To measure reproducibility, **Claude Sonnet** re-graded a ~20% overlap of **239 pairs across 11 queries**: the two agreed at Cohen's $\kappa = 0.927$ unweighted and 0.954 linearly weighted (96.7% exact, 100% within one grade). As Chapter 5 takes up, this κ measures *reproducibility between two LLMs*, not validity against a lawyer.

One further input was informal and human: throughout development, colleagues acting as proxy users tried the interface, and their verbal, formative comments shaped design decisions, most directly the move from a misleading "relevance %" to an honest "no strong matches" signal (§4.3.4). These were proxy users, not legal experts; their comments are reported as formative input only, never as survey findings.

4.1.2 Secondary Data

The corpus is secondary data: **120 majority opinions of the Supreme Court of the United States**, decided 1950–2019, retrieved from **CourtListener**, the open archive run by the Free Law Project (Free Law Project, 2024); none of its ranking is used. Opinions were ingested via the v4 API, cleaned to plain text, and chunked into **2,196 passages** (token-based, sentence-aware, ~900 tokens, ~12% overlap). Every chunk was embedded (0 NULL embeddings), and 119 of 120 cases were topic-labelled, the one omission a securities case correctly left unlabelled. The corpus is constitutional-heavy by design: the multi-label mix runs Constitutional 103, Criminal 65, Employment 21, Contracts 7, with single-digit tax, immigration and patent counts.

All metrics here are computed on this **frozen 120-case corpus**. The live corpus has since grown to 168 cases through a daily ingestion cron, but the evaluation set and its 1,276 judgements were fixed before that growth, so freezing the set keeps every comparison on identical ground.

4.1.3 Sampling Strategy

Two sampling decisions sit behind the data, both purposive rather than random. The corpus is a **purposive sample of landmark SCOTUS dockets**, since Phase 1 needed precedent dense enough to make relevance judgements meaningful, which random sampling would not deliver at n = 120. The query set is a **structured (quota) sample** across the five buckets, sized so each reports a per-bucket figure within budget. The designed survey would have recruited by purposive and snowball sampling, part of the future-validation design.

These choices buy relevance at the cost of representativeness, and the bias must be named. The corpus is one court, one jurisdiction and one era, leaning constitutional, so a result that holds here may not transfer to routine commercial disputes or Uzbek domestic law; the query set, however carefully bucketed, was written by one author and may encode that author's assumptions. None of this is fatal to a design-science evaluation, whose purpose is to demonstrate and measure an artefact rather than generalise to a population, but it bounds the findings' reach.

4.2 Analytical Techniques

4.2.1 Quantitative Analysis

The quantitative analysis turns graded pools into ranking-quality numbers, computed in Python from the qrels.jsonl judgements and per-method result lists. Four metrics are reported, each answering a specific part of the objectives. Other k-values (P@10, nDCG@5 and @20, MAP) were left uncomputed: the harness supports them, but at n = 120 they would add noise, not a distinct answer.

nDCG@10 is the headline metric: it rewards placing a strong case (grade 3) above a merely relevant one, discounts lower ranks, and normalises against the ideal ordering so scores compare across queries (Järvelin and Kekäläinen, 2002). Because legal relevance is graded, not binary, and a lawyer cares whether the strongest precedent comes first, it fits RQ1 and RQ2 better than any binary measure. The other three metrics separate the remaining qualities:

MRR@10 for how high the first relevant result lands (central to the exact bucket), **recall@10** for coverage, and **precision@5** for how clean the visible top of the list is.

Point estimates on 56 queries are noisy, so every metric carries a **95% percentile bootstrap confidence interval** (B = 10,000 resamples, seed 42). Method comparisons report the bootstrap distribution of the *difference*, and improvement is claimed only when that interval excludes zero. This discipline stops a corpus of this size being over-read, and licenses the disconfirmed result in §4.3.2.

Three further statistics serve the validity question, RQ3. Agreement between the two LLM graders is **Cohen's κ** (Cohen, 1960), which corrects raw agreement for chance. Because the grades are ordinal, calibration of the reranker's scores against the Opus judge uses **quadratic-weighted κ**, penalising a two-grade gap more than a one-grade gap. The honesty mechanism is assessed with **AUC**, the area under the ROC curve on the separation between in-domain and out-of-domain confidence scores, the standard measure of how cleanly a threshold divides two populations.

4.2.2 Qualitative Analysis

The qualitative strand actually carried out is light by design: the informal proxy-user feedback of §4.1.1, read for recurring concerns rather than formally coded. No themes, codes or frequencies are claimed, because the input was verbal and unsystematic.

The fuller qualitative method was planned, not executed. The free-text relevance rationales from the designed survey were to be analysed with **Braun and Clarke's (2006) six-phase thematic analysis**, coded **inductively** so themes emerge from what practitioners write, surfacing dimensions of relevance the metrics may miss; any divergence between coded reasoning and the nDCG ranking would bear on construct validity. Because the survey was not fielded, no coded output exists, and this analysis is named as future work, taken up in Chapter 5.

4.3 Findings

4.3.1 Demographic / Sample Profile

The unit of analysis is the corpus and query set, not a population of respondents, so Table 4.1 profiles the data the findings rest on rather than a sample of people.

Table 4.1: Corpus and query-bucket profile

Element	Value
Corpus	120 SCOTUS majority opinions, 1950–2019
Passages (chunks)	2,196 (~900 tokens each, ~12% overlap); 0 NULL embeddings
Embedding model	OpenAI text-embedding-3-small, 1536-d (HNSW, cosine)
Topic-classified	119/120 (multi-label: Constitutional 103, Criminal 65, Employment 21, Contracts 7, Tax 2, Immigration 2, Patent 1)
Data source	CourtListener API v4 (Free Law Project)
Query set	56 hand-authored queries

Element	Value
• fact_pattern	16
• paraphrase	12
• short	10
• exact	10
• cross_lingual	8 (5 ru + 3 uz)
Judged pairs (qrels)	1,276 (pool depth 20; grades 0=914, 1=240, 2=63, 3=59)
Ground-truth judge	Claude Opus (primary); Claude Sonnet validation on 239 pairs / 11 queries (κ = 0.927)
Practitioner validation	Designed, not deployed (future work; Appendices D and E)

4.3.2 Findings: Theme / Hypothesis 1

The first finding compares the four retrieval methods before reranking, and overturns an assumption already shipped to production.

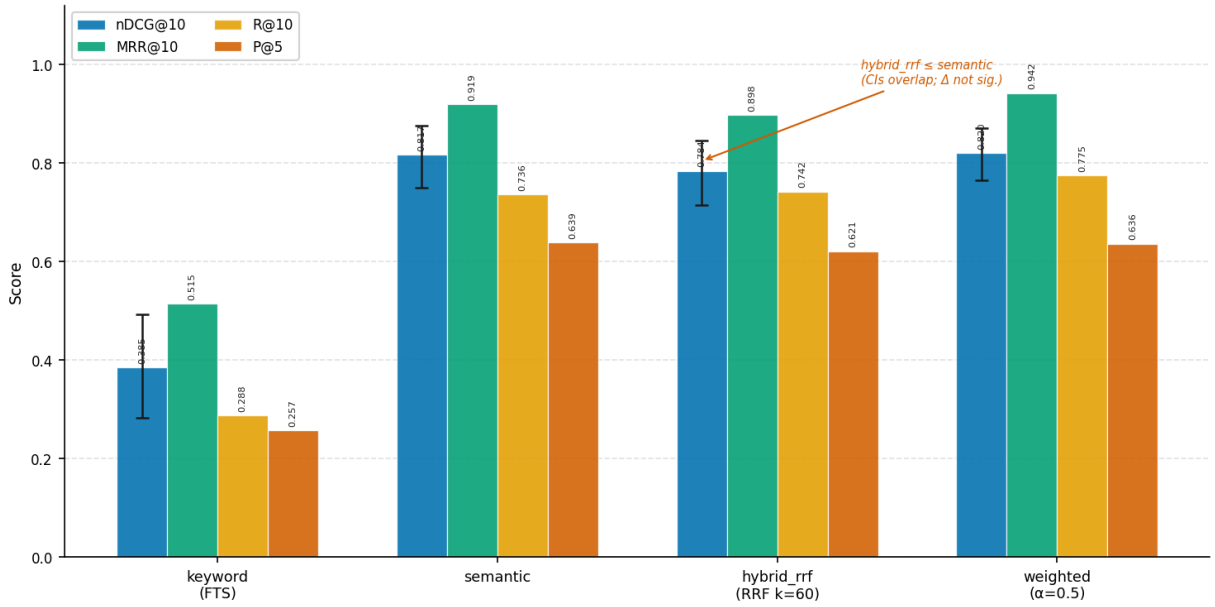


Figure 4.1: S1 baseline, retrieval metrics by method on the 120-case corpus, with 95% bootstrap CIs on nDCG@10

As Figure 4.1 shows, on the headline metric keyword (full-text) search scores nDCG@10 = 0.385 [0.283, 0.492], semantic retrieval 0.817 [0.749, 0.876], hybrid RRF (k = 60) 0.784 [0.715, 0.846] (Cormack, Clarke and Büttcher, 2009), and weighted fusion (α = 0.5) 0.820 [0.765, 0.870]. Two patterns matter. First, keyword search is **significantly the weakest**: its interval sits far below the others without overlap, and in the cross-lingual bucket it collapses to nDCG@10 = 0.000, because Cyrillic and Uzbek text does not tokenise in an English full-text index, the vocabulary-mismatch problem that motivated the project (Blair and Maron, 1985; Van Opijnen and Santos, 2017).

The second, sharper pattern is that **hybrid RRF does not beat pure semantic retrieval**: 0.784 against 0.817, intervals overlapping. The team had assumed the opposite and shipped RRF to production; the S1 evaluation disproved it (mistake M007). Weighted fusion recovers a small edge over RRF on MRR@10 (+0.044) and nDCG@10 (+0.037), the delta interval excluding zero, but does not separate from semantic either. On a 120-case corpus the dense channel does the work, and adding a lexical channel by RRF buys nothing, which turned the project from fusion tuning toward reranking, where the next finding picks up.

4.3.3 Findings: Theme / Hypothesis 2

The second finding measures whether a pointwise LLM reranker, layered on the shipped weighted-plus-exact ranking, improves quality. Unlike the first assumption, this one survives.

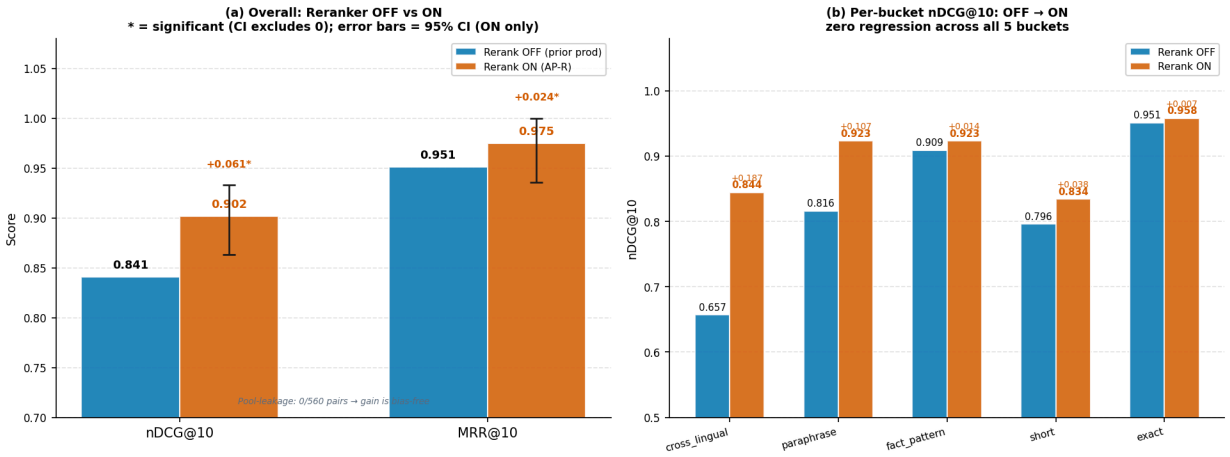


Figure 4.2: AP-R reranker OFF versus ON, overall nDCG@10 and MRR@10, and per-bucket nDCG@10

Figure 4.2 plots the comparison. With the reranker off (prior production config, edge_http), nDCG@10 = 0.841. Turning the AP-R reranker on (edge_http_rerank: gpt-4o-mini pointwise, K = 20, temperature 0) raises it to **0.902 [0.863, 0.933]** and MRR@10 to 0.975 [0.936, 1.000]. Against the strongest unranked baseline, semantic, the gain is **nDCG@10 +0.085 [0.031, 0.151]** and **MRR@10 +0.056 [0.009, 0.116]**, both intervals exclude zero, so both are statistically supported, not noise. The per-bucket breakdown shows the gain is concentrated: cross_lingual 0.657 → 0.844, paraphrase 0.816 → 0.923, short 0.796 → 0.834, fact_pattern 0.909 → 0.923, exact 0.951 → 0.958, with **zero regression** in any bucket. It is also bias-free: because K = 20 equals the pooling depth, there was **0/560 pool leakage**, so no rerank-promoted document escaped the judged pool to inflate the score. The reranker earns its place exactly where the dense channel struggled most, namely cross-lingual and paraphrased queries.

4.3.4 Findings: Theme / Hypothesis 3

The third finding concerns honesty, not ranking: can the system recognise when it has *no* strongly-relevant precedent and say so, instead of returning its least-bad guess? An early interface showed the raw fusion score as a “relevance %”, which misled proxy users; worse, against an out-of-domain query (“car accident” put to a SCOTUS corpus) it would still rank ten

cases as if they answered it. The absolute-relevance mechanism, a "no strong matches" banner raised when the top reranker score falls below a threshold of TAU = 5, closes that gap.

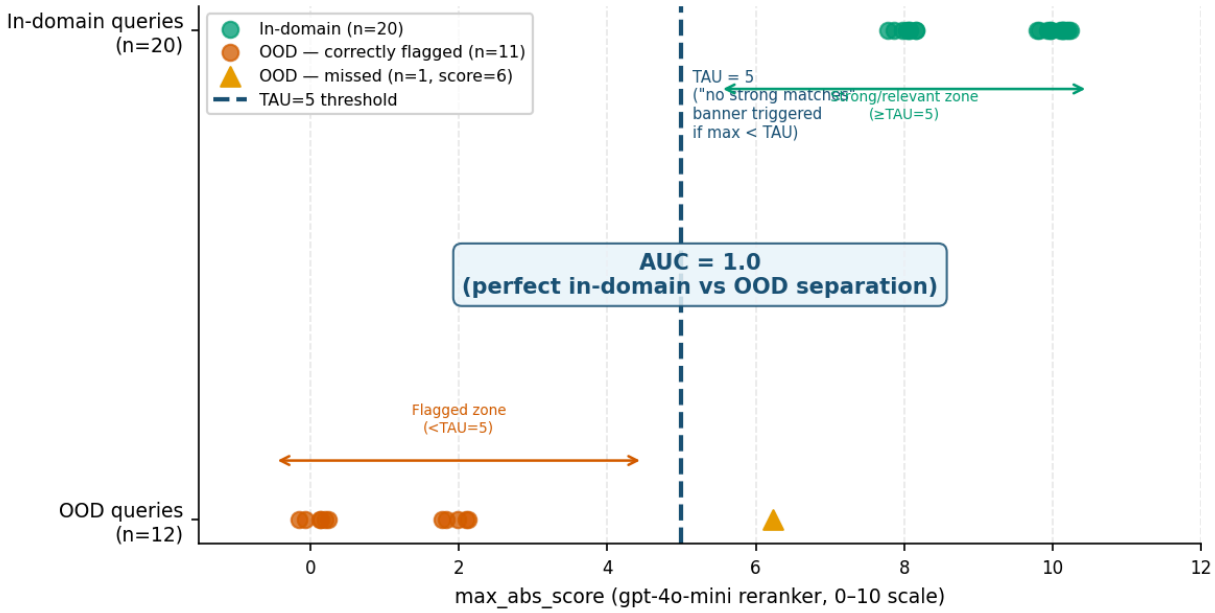


Figure 4.3: Out-of-domain honesty, in-domain versus out-of-domain score separation (AUC = 1.0, TAU = 5)

As Figure 4.3 shows, across 12 out-of-domain and 20 in-domain probes the separation between the two populations' top absolute scores is perfect: **AUC = 1.0**. The mechanism raised the "no strong matches" signal on **11 of 12** out-of-domain queries with **0/20 false suppressions** in-domain, it never denied a good precedent when one existed. The reranker's scores also calibrate reasonably against the human-standard judge: **quadratic-weighted κ vs Opus = 0.641** on 1,119 scored-and-judged pairs (exact agreement 0.613). One result is reported against the project's interest: a cheaper gpt-4o-mini judge tested as an Opus substitute reached only **κ = 0.716 quadratic-weighted**, below the 0.75 bar, so it is a *marginal* judge and Opus was kept as ground truth. Whether the AUC = 1.0 separation and κ = 0.641 calibration match what a practising lawyer calls "no good match" is the open question for the practitioner validation named in §4.2.2.

4.4 Comparison of Patterns and Trends

The three findings tell one coherent story. The dense semantic channel carries the quality, lifting nDCG@10 from 0.385 to 0.817 alone, while the lexical channel RRF adds does not move the metric on this corpus and cannot contribute at all cross-lingually. The reranker adds a further, statistically supported +0.085 over semantic, but unevenly: largest where retrieval is hardest (cross_lingual +0.187, paraphrase +0.107) and smallest where it was already near-perfect (exact +0.007). The honesty mechanism adds perfect out-of-domain separation (AUC = 1.0) with no in-domain false suppression, trading nothing measurable for that safety. The metrics improve fastest on the buckets a single aggregate would have ignored, which vindicates per-bucket reporting.

These conclusions are drawn entirely from the quantitative strand against an LLM judge, the limit the comparison cannot yet cross: whether a practitioner agrees the reranked list and the "no strong matches" verdict are right awaits the designed-but-not-deployed survey named as future work. The picture also rests on a 120-case, single-court corpus whose headline numbers are scheduled for re-measurement at scale. Chapter 5 takes up how far this evidence can be trusted.

Chapter 5: Discussion

5.1 Interpretation of Findings

RQ1 asked how far semantic and hybrid retrieval improve relevance over keyword search. For the dense channel the answer is yes: semantic retrieval roughly doubles keyword's ranking quality. For fusion it is no: the team shipped Reciprocal Rank Fusion assuming a hybrid beats its parts, but S1 disproved that (mistake M007), since on a small, semantically uniform corpus the dense channel holds almost all the recoverable signal.

RQ2 asked whether a pointwise LLM reranker improves ranking further, and whether it can admit when no strong precedent exists. Both resolve in the system's favour. Its lift over semantic (about 0.085 of nDCG@10, the interval clear of zero) falls almost entirely on the cross-lingual and paraphrase buckets, as a second-stage model should. The more consequential result is the honesty mechanism: a pure ranker always returns a top-ten list even when nothing fits, a real hazard in law. Here the absolute scores told in-domain from out-of-domain perfectly (AUC 1.0), flagging all but one of the twelve out-of-domain queries and suppressing no genuine result.

RQ3 matters most: it asks not what the system scores but whether the score can be believed. The LLM judge is highly reproducible, but reproducibility is agreement between judges of the same kind and says nothing about whether either tracks a practising lawyer's sense of relevance; a cheaper model judge agreed only marginally, so "an LLM judge" varies by model. Chapter 4's relative comparisons are therefore trustworthy, being reproducible and blind, while the absolute scores measure relevance as a capable model construes it, not as a lawyer would. That distinction governs this chapter.

5.2 Comparison with the Literature

On RQ1 the results sit inside the literature. The margin of semantic over keyword search reproduces Karpukhin et al. (2020), whose dense passage retriever beat strong lexical baselines, and falls where dense retrieval is strongest: a single, homogeneous domain. The lexical floor is the one Blair and Maron (1985) quantified.

Fusion seems to contradict Cormack, Clarke and Büttcher (2009), who established that Reciprocal Rank Fusion reliably beats its input rankings, but it does not: RRF gains most when its rankers contribute independent evidence, and on 120 constitutional-heavy opinions the lexical and dense rankers largely agree, while the English full-text channel adds nothing cross-lingually. With little independent signal to fuse, the result fits the BEIR warning that a hybrid does not always win (Thakur et al., 2021); the lexical channel's weakness is structural, not a tuning failure (Robertson and Zaragoza, 2009).

The reranking findings agree with the literature. Nogueira and Cho (2019) established that a second-stage neural reranker can substantially lift a first-stage list, and Sun et al. (2023) showed a general-purpose LLM can rerank competitively with no task-specific training, which the pointwise reranker reproduces on a legal corpus it was never trained for. The RQ3 caution is the one Zheng et al. (2023) raised: an LLM judge can reach high agreement and still carry

shared, systematic biases. The κ near 0.93 here, agreement within one model family, is no stronger an endorsement than they offer.

5.3 Validity

Validity is examined along three axes, and they do not reach the same verdict.

Construct validity is the weakest link, weak by design rather than accident. Every nDCG@10 figure rests on relevance grades assigned by an LLM, so the metric measures the judge's construct of relevance, not legal correctness. Replacing the planned student hand-labelling with an LLM judge (mistake M009) shifted that construct, and it bites hardest on the absolute scores: a reported nDCG@10 of 0.902 states the judge's relevance model, not legal fact. This is the single most serious threat, and why the relative, blinded comparisons are treated as load-bearing and the absolute numbers as provisional. The matched-snippet headline offsets it a little, since the card a user reads is built from retrieved text, not a generated summary, but no design choice closes the gap; only practitioner labels can, named here as future work.

Internal validity is comparatively strong, carried by three deliberate choices. Candidates were pooled at depth 20 in the TREC style with the producing method hidden from the judge, so no method was rewarded for being recognisable. Case relevance is scored from the best-matching passage, not the average, so one strong passage is not diluted by weaker ones. Most important for the reranker claim, the reranker ran at K = 20, equal to the pooling depth, giving zero pool leakage across all 560 reranked pairs. Because no reranked document escaped the judged pool, the rerank-on-versus-off gain cannot be an artefact of conveniently unjudged documents.

External validity is bounded. The corpus is 120 majority opinions from one court, ending in 2019, leaning heavily towards constitutional and criminal matters. Generalisation to other courts, to fact-heavy lower-court litigation, to other jurisdictions, and above all to the trilingual use case the interface advertises is unproven. The cross-lingual evidence is especially thin at eight queries, so the strong reranker lift there is a signal worth chasing, not a settled result. None of this voids the internal comparisons, but it confines them to the corpus they were measured on, which is why the headline is scheduled for re-measurement at scale.

5.4 Reliability

Reliability asks whether a repeat of the study would land in the same place, and here the evaluation stands on firmer ground than its validity. The inter-judge agreement near κ 0.93, with complete agreement to within one grade on the overlap sample, shows the labels are reproducible: a second judge of the same class rebuilds them almost exactly. The inference is just as repeatable, because every confidence interval came from a percentile bootstrap with a fixed configuration (10,000 resamples, seed 42), so re-running the harness returns identical intervals, not fresh random ones. That determinism is what makes a small study auditable.

Reliability is not freedom from systematic error. One qualification is ordinary sampling noise: with 56 queries across five buckets, the per-bucket intervals are wide, so the aggregate deltas survive (their intervals exclude zero) while the bucket-level story is suggestive, not precise. The deeper qualification, and why §5.3 holds the binding constraint on confidence, is that the ground truth rests on a single family of judge models: high agreement between two of its

members shows consistency within that family but cannot detect a bias shared across it. If such models systematically over-weight topical similarity against doctrinal relevance, a repeat run would reproduce that bias faithfully and report it as reliable. High reliability and a real validity gap can therefore coexist, as here.

5.5 Project Effectiveness Evaluation

The four objectives are assessed on a red–amber–green basis in Table 5.1. Three are complete; the fourth is rated against its agreed scope, not the practitioner survey that was deliberately dropped.

Table 5.1: Project effectiveness evaluation by objective (RAG)

Objective	RAG	Evidence and comment
O1: Critically review the literature to locate the gap in legal IR, dense/hybrid retrieval and reranking	**Green**	Chapter 2 reviews legal IR, dense retrieval, fusion and reranking, and rejects named alternatives (D1).
O2: Design and implement the hybrid pipeline with an LLM reranker and an honest no-match signal	**Green**	The shipped pipeline (text-embedding-3-small, weighted fusion with exact-match prepend, AP-R reranker at K = 20, matched-snippet headline) runs on the 120-case corpus, verified by browser smoke test after M003.
O3: Build a reproducible evaluation harness and measure quality across query types	**Green**	TREC-style pooling, graded 0–3, LLM-as-judge (κ near 0.93), bootstrap CIs and per-bucket metrics produced the S1, AP1 and AP-R results; headline re-measurement deferred to ~1,000 cases.
O4: Critically evaluate the evidence and recommend, characterising practitioner judgement from the literature	**Green**	Validity and reliability are evaluated in §5.3–5.4 and practitioner relevance is characterised from the literature (Van Opijnen and Santos, 2017; Zheng et al., 2023). Primary practitioner validation was scoped out as future work, not a deliverable; the designed instrument (Appendices D and E) evidences that planning. Achieved as scoped, with the human-validation gap stated openly in §5.3 and §5.6.

5.6 Limitations

Five limitations bear materially on the findings, each stated with its effect.

- **Small, homogeneous corpus.** 120 single-court, constitutional-heavy opinions are too few for a production-scale claim, and the likeliest reason fusion added nothing (§5.1); every absolute number is therefore provisional, the headline deferred to roughly a thousand cases.

- **LLM judge, not human ground truth.** Because the labels come from a model (M009), the metrics measure machine-construed relevance; the absolute scores cannot yet be read as legal correctness, and only the relative, blinded comparisons are firmly defensible. Practitioner labelling is future work.
- **Thin cross-lingual evidence.** The cross-lingual bucket rests on eight queries with the translation step still flag-gated, so the strong reranker lift there is a signal on a tiny sample, not a validated multilingual result.
- **Operating cost and latency unmeasured.** Query-time latency and the per-search cost of twenty reranker calls were never measured, so feasibility at scale rests on quality alone, not economics.
- **Human triangulation absent by design.** No practitioner data exists, because the survey was cancelled rather than postponed, so the project cannot say how far experts agree with the reranked ordering or the no-match verdict; for human validity, RQ3 rests only on the literature.

5.7 Implications for Practice and Future Research

The most transferable contribution is not the reranker's accuracy gain but the honest no-match signal. Returning a best guess to every query is dangerous in law: a "car accident" query put to a constitutional-law corpus should return nothing, not a plausible but doctrinally wrong precedent. Telling in-domain from out-of-domain queries at no measurable cost to genuine results points to a design pattern for safety-critical retrieval: report a ranking with a calibrated confidence that it is worth trusting, and abstain when it is not. The implication is to invest in the reranker and the abstention signal rather than further fusion tuning, which buys nothing on a corpus of this kind. The commercial market is already moving this way as it rebuilds legal research around generative models (Thomson Reuters, 2023, 2024), though those tools keep their effectiveness claims private where this project measures them in the open.

Four research directions follow. Scale comes first: re-running the comparison on roughly a thousand cases would test whether the dense-dominates-fusion pattern survives a larger, less homogeneous corpus, and the mechanism is already live, a daily ingestion job adding about thirty cases a day. Second, and most limiting to current confidence, is human practitioner labelling and recomputation of the metrics against it, the only route from a reproducible evaluation to a valid one. Third is a genuine multilingual evaluation, with the translation step enabled and far more than eight cross-lingual queries. Fourth is operational: query-time latency and per-search cost, since a quality gain a system cannot afford to serve is no gain. Beyond the project, the family-shared-bias problem of §5.4 reaches past legal retrieval: for any LLM-as-judge evaluation, inter-judge κ is necessary but not sufficient.

Chapter 6: Conclusion and Recommendations

6.1 Summary of the Project

This project designed and critically evaluated a relevance-ranked retrieval system that recommends similar judicial precedents from a plain-language legal description. It targeted vocabulary mismatch: lexical search returns shared words, not shared legal issues. Following a design-science logic, the work built and interrogated a real artefact: a hybrid pipeline over a 120-case SCOTUS corpus, using OpenAI `text-embedding-3-small` embeddings, an optional pointwise `gpt-4o-mini` reranker, and an honest "no strong matches" signal. The instrument was a reproducible harness: 56 queries across five buckets, 1,276 graded pairs, an LLM-as-judge, and a 95% bootstrap confidence interval on every difference.

Several findings went against expectation. Semantic retrieval beat keyword search decisively (nDCG@10 0.385 to 0.817), yet on non-English queries the lexical channel collapsed to zero. Against the team's prior assumption, hybrid Reciprocal Rank Fusion did not beat the dense channel (0.784 against 0.817); that belief was already in production, and only the evaluation caught it. The pointwise reranker survived: +0.085 nDCG@10 over the semantic baseline, the interval excluding zero, and the cross-lingual bucket lifted from 0.657 to 0.844. The honesty mechanism separated in-domain from out-of-domain queries perfectly (AUC = 1.0) and suppressed no genuine in-domain match.

6.2 Achievement of Objectives

- **O1 (review): met.** Chapter 2 critically reviewed the literature on legal information retrieval, dense and semantic search, hybrid ranking and reranking; alternatives were weighed and rejected, and the gap was located precisely, not asserted.
- **O2 (design and implement): met.** A working hybrid pipeline was built over the 120-case corpus: weighted fusion with a deterministic exact-match channel, the AP-R reranker at K = 20, and the "no strong matches" banner, shipped behind the `court-case-search` edge function and verified live.
- **O3 (evaluate): met.** The reproducible harness measured quality across five query types with per-bucket breakdowns and bootstrap confidence intervals, every ranking change gated on the eval. The headline at scale was deferred to a larger corpus.
- **O4 (validate and recommend): met as scoped.** Validity and reliability were evaluated (§5.3–5.4): the reranker was calibrated against the Opus reference judge (quadratic-weighted $\kappa = 0.641$), and practitioner judgement was characterised from the literature. Primary practitioner validation, collecting human relevance labels from qualified lawyers, was deliberately left to future work. The report is candid about the cost: RQ3 is answered for the labels' reproducibility but only partially for their validity against a practising lawyer, and closing it is the first recommendation below.

6.3 Contribution to Knowledge and Practice

The contribution to knowledge is a reproducible, confidence-interval-reported account of where each pipeline layer earns its keep in issue-level legal relevance on a fixed corpus. Two results stand out. First, hybrid fusion need not beat a strong dense channel: on a homogeneous

single-court corpus RRF added nothing, a counterweight to the assumption that fusion is always an improvement. Second, the project calibrates an LLM relevance judge for legal text, separating what it measures well (reproducibility, $\kappa \approx 0.93$ between two graders) from what it cannot yet establish (validity against a human lawyer); a cheaper judge fell short of its bar ($\kappa = 0.716$, below 0.75).

In practice, the honest no-match signal is a safety feature, not cosmetic: it turns confident-but-wrong behaviour on out-of-domain queries into an explicit refusal, which matters more in a legal tool than a marginal ranking gain. The eval-first, eval-gated discipline transfers beyond law: build the measuring instrument before optimising it.

6.4 Recommendations

6.4.1 For Practitioners / Industry

- Adopt an absolute-relevance score with an honest no-match threshold before scaling a legal retrieval tool, so it can decline rather than fabricate a ranking it cannot stand behind.
- Validate retrieval against a held-out evaluation set with confidence intervals before shipping, and report quality per query type, not a single aggregate. The RRF default was wrong in production because it had never been measured; the layers that help most often help only on segments an average hides.
- Treat a strong dense baseline as the bar any fusion layer must beat before it ships, rather than assume fusion helps.

6.4.2 For Future Researchers

- Collect human relevance labels from qualified practitioners and recompute the metrics against them; this is the most valuable next step, the only way to turn a reproducible LLM evaluation into a valid one. The survey instrument (Appendices D and E) is ready to field.
- Repeat the evaluation on a larger, multi-court corpus, since 120 single-court cases are too homogeneous to settle the per-bucket questions or fusion-versus-dense at scale.
- Extend the work to multilingual legal IR, with the translation step enabled and far more than eight cross-lingual queries, so the strong reranker lift can be confirmed, not merely indicated.
- Quantify the latency and per-query cost of the LLM reranker, so the ranking gain can be weighed against serving cost.

6.5 Closing Remarks

The clearest lesson is that a retrieval system and the evidence for it are two different things, and the second is easier to overestimate. The most instructive moments were the disproved hybrid assumption, the cheap judge that proved only marginal, and the choice to leave human validation for future work rather than claim it. A system that knows when it has no good answer, and a project that knows how far its evidence reaches, are the two honest results worth keeping.

Chapter 7: Reflective Evaluation of Personal and Professional Development

7.1 Choice of Reflective Model

I chose Gibbs' (1988) reflective cycle to structure this chapter. Its six stages run from describing the event, through feelings and evaluation, into analysis, a concrete conclusion and an action plan, which fits a look back at a finished capstone better than the alternatives. I rejected Kolb's (1984) experiential cycle as too abstract and continuous for a single, bounded project like this one. I also set aside Schön's (1983) reflection-in-action, because it describes thinking on your feet during the work, whereas I am reflecting after it, with the evidence in front of me.

7.2 Reflection Using Gibbs' Cycle

7.2.1 Description: What happened?

Working alone, I designed and built court-cases, a relevance-ranked search over a CourtListener corpus of United States Supreme Court opinions (120 cases frozen for evaluation, now 168 live). I directed an agentic-coding workflow (Claude Code, an AI coding agent) to build the pipeline: OpenAI embeddings, Reciprocal Rank Fusion (RRF), then an LLM reranker. I ran a 56-query, 1,276-judgement evaluation with an LLM-as-judge, shipped the AP-R reranker as the production ranker, and deployed the search edge function and a daily cron on Railway that keeps the corpus growing.

7.2.2 Feelings: What were you thinking and feeling?

My strongest feelings clustered around being wrong on the record, even when the only audience was that record. The S1 evaluation showed that RRF hybrid did not beat semantic retrieval. That was the method I had already shipped. I felt deflated for an afternoon, then oddly clear: the measurement had done its job (M007). I felt a sharper embarrassment over M001, where I had declared the court tables empty from a missing checkpoint file when the database held 120 opinions all along. The hidden CourtListener daily quota humbled me too, turning a confident two-to-three-day scaling plan into roughly twenty-five days (M018). Finding a Supabase token sitting in git history was a colder kind of jolt (M012). Against all that, the reranker holding up at evaluation, and the honest no-strong-matches banner behaving exactly as I had designed it, felt good, and earned.

7.2.3 Evaluation: What was good and bad?

What went well was discipline around evidence. No ranking change reached production without clearing the eval-gate, so a metric, not a hunch, decided what shipped. The mistakes log earned its keep: at M010 the habit it had trained, checking both retrieval channels, caught a corpus-polluting bug before any ingest ran. I also reported the weak parts honestly: the inter-judge κ, the bootstrap intervals, the concession that an LLM judge is not human ground truth. What went badly was timing and hygiene. I found hard external limits too late, as the daily quota showed (M018). I kept reasoning from whatever environment was convenient rather than one that mirrored production, which broke a deploy plan, a scale-ingest run and a minimal cron

image in turn (M013, M016, M021). My version-control discipline slipped too: a leaked secret (M012), and parallel sessions racing on one checkout (M020).

7.2.4 Analysis: Why did it happen this way?

Most of these failures share one root: I trusted a description of reality instead of checking reality. Working alone, with no colleague to challenge a claim, status-by-proxy crept in, and a missing file stood in for a database I never queried (M001). My ranking design started from a belief I had read, that hybrid beats semantic, rather than a measurement on this corpus, and the evaluation corrected me (M007). The same habit ran through the build: I treated my own plans as true until a runtime check in a production-like environment disproved them (M013, M016, M021). Gibbs (1988) frames why setting this down matters: learning comes from reflecting on concrete experience, not merely from living through it. The same caution applied to the tooling. The agentic-coding workflow was productive, but only safe because I verified its output against the eval-gate and the tests rather than trusting it.

7.2.5 Conclusion: What did you learn?

Four lessons survive the project. First, I verify state at its source before I assert it: a count query or a live probe, never an artefact standing in for the thing itself. Second, I treat my own plans with the same scepticism as external or AI advice, because both are hypotheses to be tested against a runtime that mirrors production, not facts to build on. Third, an empty or NULL field is a question, not an answer. I almost accepted forty-six cases as undecided because one column was blank, when they were decided and simply unmapped (M023). Fourth, a deviation from the plan belongs in the methodology when it happens, not absorbed in silence. When I swapped manual relevance labels for an LLM judge, I should have recorded the change and its effect on what I can claim (M009).

7.2.6 Action Plan: What will you do differently?

In future projects I will:

- run a source-of-truth check, a query or a live probe, before making any claim about system or data state;
- pre-mortem every external dependency, reading all of its rate-limit tiers, before I build a timeline on it;
- run a dry-run in a production-mirror environment before any long or irreversible ingest or deploy;
- scan each commit’s diff for secrets before I push.

7.3 SWOT Analysis of Personal Development

Table 7.1 summarises the position.

Table 7.1: SWOT analysis of personal development

Strengths	**Weaknesses**
• Directing and verifying agentic-coding tooling (Claude Code): accepting, rejecting and	• Solo confirmation bias, no challenger to a claim (M001, M007).• Under-scoped

Strengths	Weaknesses
re-running output, gated on the eval.• Built an end-to-end IR/eval pipeline (embeddings, RRF, reranker, nDCG/MRR, bootstrap CIs).• Intellectual honesty: the mistakes log; the LLM judge conceded as not human truth.• Design-science discipline: nothing shipped without the eval-gate.	external constraints (M018 quota; M009 annotation cost).• Over-trusted my plans over a runtime check (M013, M016, M021).• Hygiene lapses: a leaked token, parallel-session races (M012, M020).
Opportunities	**Threats**
• Scaling court-cases toward ~1,000 cases with practitioner validation.• Growing the legal/regulatory-technology venture JurisBase anchors.• An under-served Central-Asian legaltech market (DataReportal, 2024).	• Solo-founder bandwidth.• Dependence on external API quotas (CourtListener, OpenAI).• A fast-moving AI-IR field that can outdate the approach.

7.4 Transferable Skills Gained

Table 7.2 maps these skills to evidence.

Table 7.2: Transferable skills gained (self-assessed 1–5, with project evidence)

Skill	Level	Evidence from this project
Directing and verifying agentic-coding tooling	4	Directed Claude Code to build the pipeline and harness, then accepted, rejected and re-ran its output and gated every change on the eval. Directing and verifying such tooling is an emerging professional capability (Ziegler et al., 2024).
Information-retrieval engineering and evaluation	4	Built embeddings, RRF, weighted fusion and the AP-R reranker, and measured them with nDCG@10, MRR, recall and bootstrap CIs on the 120-case corpus.
Critical thinking and scientific scepticism	4	Disproved my own assumption that RRF beats semantic retrieval on this corpus, and let the measurement redirect the design (M007).
Data discipline and source-of-truth verification	3	Diagnosed forty-six NULL decision-date rows as a mapping gap, not undecided cases, by checking the source and content (M023).
Solo project management	3	Ran the work to a WBS and Gantt, using the mistakes log (M001–M023) and the eval-gates as the tracking system.
Resilience	4	Recovered a broken production cron after a wrong first diagnosis and a guard that froze ingestion (M022, M023).

7.5 Future Development Plan

Reviewing where this leaves me, the honest gap is between a working artefact and a validated one. court-cases is a real system I can demonstrate, not a paper design, but its relevance numbers still rest on an LLM judge, and the practitioner study I designed was never run. My plan closes that gap and grows the venture the system anchors. It also settles a strategic choice I have not yet made, deciding it from evidence, not preference (Table 7.3).

Table 7.3: Future development plan

Goal	Action	Timeframe
Short term: take court-cases from the 120-case evaluation corpus toward ~1,000 cases and run the practitioner validation I designed but did not deploy.	Scale the live corpus through the daily cron; recruit a few practising lawyers, study how they judge relevance, and measure the tool's value to them.	3–6 months
Medium term: use court-cases as the flagship piece of my portfolio when applying to top universities for a Master's.	Present it as a demonstrable system, not just a report; strengthen the study and the agent-directed evaluation tooling behind it.	1–2 years
Long term: grow JurisBase as a foundation of my wider venture and settle its model.	Weigh a real strategic choice: a paid product in my own ecosystem, or an open, Wikipedia-style free legal-knowledge resource that builds public value and a personal brand.	3–5 years

Reference List

1. American Bar Association (2023) *ABA TechReport 2023*. Chicago: American Bar Association Law Practice Division. Available at: https://www.americanbar.org/groups/law_practice/resources/tech-report/2023/ (Accessed: 7 June 2026).
2. Blair, D.C. and Maron, M.E. (1985) ‘An evaluation of retrieval effectiveness for a full-text document-retrieval system’, *Communications of the ACM*, 28(3), pp. 289–299. doi:10.1145/3166.3197.
3. Braun, V. and Clarke, V. (2006) ‘Using thematic analysis in psychology’, *Qualitative Research in Psychology*, 3(2), pp. 77–101. doi:10.1191/1478088706qp063oa.
4. Buckley, C. and Voorhees, E.M. (2000) ‘Evaluating evaluation measure stability’, in *Proceedings of the 23rd Annual International ACM SIGIR Conference*, Athens, July 2000. New York: ACM, pp. 33–40. doi:10.1145/345508.345543.
5. Cohen, J. (1960) ‘A coefficient of agreement for nominal scales’, *Educational and Psychological Measurement*, 20(1), pp. 37–46. doi:10.1177/001316446002000104.
6. Cormack, G.V., Clarke, C.L.A. and Büttcher, S. (2009) ‘Reciprocal rank fusion outperforms condorcet and individual rank learning methods’, in *Proceedings of the 32nd Annual International ACM SIGIR Conference*, Boston, July 2009. New York: ACM, pp. 758–759. doi:10.1145/1571941.1572114.
7. Creswell, J.W. (2014) *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. 4th edn. Thousand Oaks, CA: SAGE.
8. DataReportal (2024) *Digital 2024: Uzbekistan*. DataReportal / We Are Social / Meltwater, 23 February. Available at: https://datareportal.com/reports/digital-2024-uzbekistan (Accessed: 7 June 2026).
9. Free Law Project (2024) *About Free Law Project*. Oakland: Free Law Project. Available at: https://free.law/about/ (Accessed: 7 June 2026).
10. Gibbs, G. (1988) *Learning by Doing: A Guide to Teaching and Learning Methods*. Oxford: Further Education Unit, Oxford Polytechnic.
11. Hevner, A.R., March, S.T., Park, J. and Ram, S. (2004) ‘Design science in information systems research’, *MIS Quarterly*, 28(1), pp. 75–105.
12. Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A. and Grave, E. (2022) ‘Unsupervised dense information retrieval with contrastive learning’, *Transactions on Machine Learning Research*. arXiv:2112.09118. Available at: https://arxiv.org/abs/2112.09118 (Accessed: 7 June 2026).
13. Järvelin, K. and Kekäläinen, J. (2002) ‘Cumulated gain-based evaluation of IR techniques’, *ACM Transactions on Information Systems*, 20(4), pp. 422–446. doi:10.1145/582415.582418.
14. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W. (2020) ‘Dense passage retrieval for open-domain question answering’, in *Proceedings of EMNLP 2020*, Online, November 2020. Stroudsburg, PA: ACL, pp. 6769–6781. doi:10.18653/v1/2020.emnlp-main.550.
15. Kolb, D.A. (1984) *Experiential Learning: Experience as the Source of Learning and Development*. Englewood Cliffs, NJ: Prentice-Hall.

16. Muennighoff, N., Tazi, N., Magne, L. and Reimers, N. (2023) ‘MTEB: Massive text embedding benchmark’, in *Proceedings of EACL 2023*, Dubrovnik, May 2023. Stroudsburg, PA: ACL, pp. 2014–2037. doi:10.18653/v1/2023.eacl-main.148.
17. Nogueira, R. and Cho, K. (2019) *Passage re-ranking with BERT*. arXiv:1901.04085. Available at: https://arxiv.org/abs/1901.04085 (Accessed: 7 June 2026).
18. Nogueira, R., Jiang, Z., Pradeep, R. and Lin, J. (2020) ‘Document ranking with a pretrained sequence-to-sequence model’, in *Findings of the ACL: EMNLP 2020*, Online, November 2020. Stroudsburg, PA: ACL, pp. 708–718. doi:10.18653/v1/2020.findings-emnlp.63.
19. OpenAI (2024) ‘New embedding models and API updates’, *OpenAI Blog*, 25 January. Available at: https://openai.com/index/new-embedding-models-and-api-updates/ (Accessed: 7 June 2026).
20. Peffers, K., Tuunanen, T., Rothenberger, M.A. and Chatterjee, S. (2007) ‘A design science research methodology for information systems research’, *Journal of Management Information Systems*, 24(3), pp. 45–77. doi:10.2753/MIS0742-1222240302.
21. Reimers, N. and Gurevych, I. (2019) ‘Sentence-BERT: sentence embeddings using Siamese BERT-networks’, in *Proceedings of EMNLP-IJCNLP 2019*, Hong Kong, November 2019. Stroudsburg, PA: ACL, pp. 3982–3992. doi:10.18653/v1/D19-1410.
22. Robertson, S. and Zaragoza, H. (2009) ‘The probabilistic relevance framework: BM25 and beyond’, *Foundations and Trends in Information Retrieval*, 3(4), pp. 333–389. doi:10.1561/1500000019.
23. Saunders, M.N.K., Lewis, P. and Thornhill, A. (2023) *Research Methods for Business Students*. 9th edn. Harlow: Pearson.
24. Schafer, B. and Maxwell, T. (2010) ‘Natural language processing and query expansion in legal information retrieval: challenges and a response’, *International Review of Law, Computers and Technology*, 24(1), pp. 63–72. doi:10.1080/13600860903570194.
25. Schön, D.A. (1983) *The Reflective Practitioner: How Professionals Think in Action*. New York: Basic Books.
26. Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D. and Ren, Z. (2023) ‘Is ChatGPT good at search? Investigating large language models as re-ranking agents’, in *Proceedings of EMNLP 2023*, Singapore, December 2023. Stroudsburg, PA: ACL, pp. 14918–14937. doi:10.18653/v1/2023.emnlp-main.923.
27. Susskind, R. (2023) *Tomorrow’s Lawyers: An Introduction to Your Future*. 3rd edn. Oxford: Oxford University Press.
28. Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. and Gurevych, I. (2021) ‘BEIR: a heterogenous benchmark for zero-shot evaluation of information retrieval models’, in *Proceedings of NeurIPS 2021, Datasets and Benchmarks Track*, Online, December 2021. arXiv:2104.08663. Available at: https://arxiv.org/abs/2104.08663 (Accessed: 7 June 2026).
29. Thomson Reuters (2023) ‘Thomson Reuters completes acquisition of Casetext, Inc.’, *Press Release*, 17 August. Available at: https://www.thomsonreuters.com/en/press-releases/2023/august/thomson-reuters-completes-acquisition-of-casetext-inc (Accessed: 7 June 2026).
30. Thomson Reuters (2024) *Future of Professionals Report 2024*. Thomson Reuters, July. Available at:

https://www.thomsonreuters.com/en-us/posts/technology/future-of-professionals-2024/ (Accessed: 7 June 2026).
31. Van Opijnen, M. and Santos, C. (2017) ‘On the concept of relevance in legal information retrieval’, *Artificial Intelligence and Law*, 25(1), pp. 65–87. doi:10.1007/s10506-017-9195-8.
32. Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E. and Stoica, I. (2023) ‘Judging LLM-as-a-judge with MT-Bench and Chatbot Arena’, in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, New Orleans, December 2023. Red Hook, NY: Curran Associates, pp. 46595–46623. Available at: https://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html (Accessed: 7 June 2026).
33. Ziegler, A., Kalliamvakou, E., Li, X.A., Rice, A., Rifkin, D., Simister, S., Sittampalam, G. and Aftandilian, E. (2024) ‘Measuring GitHub Copilot’s impact on productivity’, *Communications of the ACM*, 67(3), pp. 54–63. doi:10.1145/3633453.

Appendices

Appendices support but are not essential to the main argument; each is referenced from the relevant chapter.

Appendix A: Project Proposal (approved version)

Working title. Honest Precedent: designing and evaluating a relevance-ranked retrieval system for plain-language legal case search.

Background and problem. Legal precedent is still located largely by keyword and citation search, which return documents that share words rather than a legal issue. The proposal targets this vocabulary-mismatch problem within the court-cases feature of JurisBase, alongside a second problem: retrieval systems present a confident ranking even for queries with no on-point answer.

Aim and objectives. To design and critically evaluate a relevance-ranked precedent-retrieval system from plain-language descriptions, and assess its evaluation's validity against expert practice. The four objectives (locate the literature gap; build the hybrid pipeline with reranker and honest no-match signal; build a reproducible evaluation harness; evaluate validity and recommend) and the three research questions are set out in full in §1.3–§1.5.

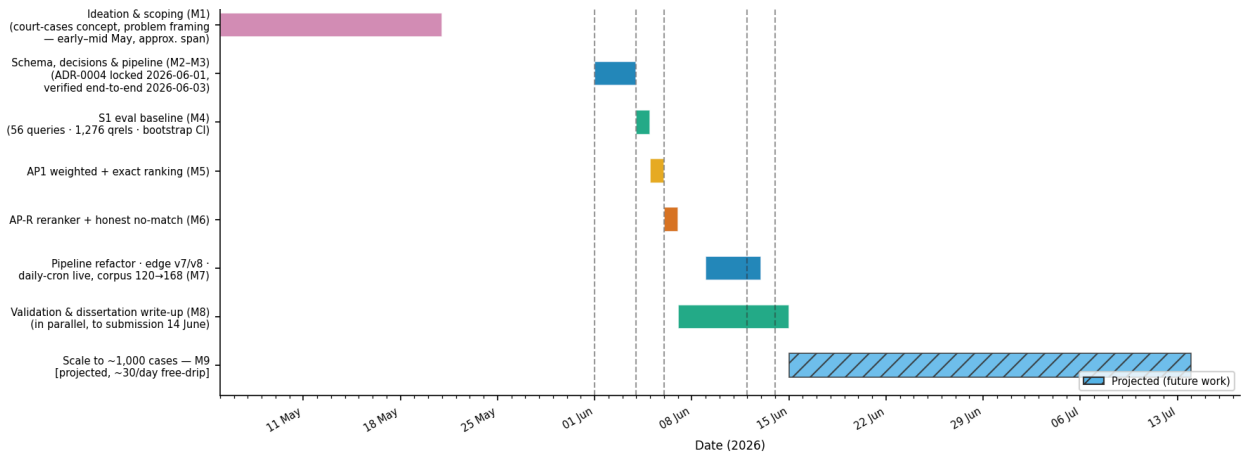
Methodology. A design-science strategy under a pragmatist philosophy, mixed-methods (quantitative retrieval metrics plus a lighter qualitative strand), managed Agile under an eval-gate on every ranking change (Ch3).

Scope. In: the court_* retrieval and evaluation work on a SCOTUS corpus. Out: the separate JurisBase law pipeline, learning-to-rank, GPU rerankers and any AI-generated cases.

Note: the O4 survey strand was re-scoped to future work (§3.7); the instrument is in Appendices D and E. Approved by supervisor on [date, to be confirmed by the author].

Appendix B: Full Gantt Chart

The full-resolution timeline is reproduced below; it appears reduced as Figure 3.2 (§3.4.2).



Appendix B: full project Gantt chart

Appendix C: Risk Register (full version)

The risk register is reproduced in full in the main text as Table 3.3 (§3.6); no separate version is needed. Likelihood and impact are scored 1–5; the residual RAG reflects the position after mitigation.

Appendix D: Ethics Approval and Consent Forms

No human-subject data was collected, so no ethics-committee approval was required (§3.7). The consent form below was **designed** for the future practitioner survey and **not deployed**; it is held as evidence of planning rigour.

Study title: Evaluating the relevance of an AI-assisted legal-precedent search system against practitioner judgement. **Researcher:** Turdiyev Abduraxmon (BTEC Level 6 Unit 2, PDP University). **Supervisor:** Abdulaziz Gulomov. Contact details are added at deployment.

What you will do. Complete one short online questionnaire (about 15–20 minutes): rate result cases on a 0–3 relevance scale, give brief reasons, and answer a few background questions. No technical knowledge is needed.

Voluntary participation, anonymity and GDPR. Participation is voluntary, with the right to stop at any time or withdraw within two weeks by quoting a participant code. No identifying information is stored; each response carries only a non-identifying code, is processed on the basis of consent, held on access-controlled storage for up to 24 months, then securely deleted.

Consent. By proceeding you confirm you have read this information and consent to your anonymised responses, including brief comments, being used and quoted in this academic project.

Appendix E: Data Collection Instruments

The practitioner relevance survey instrument **designed** as the external validity check for RQ3 (§4.2.2), **not deployed** within the project window. Result lists are **blinded** (randomised "List A"/"List B", AI scores hidden), so respondents cannot tell the baseline from the reranked system.

Section 1: About you (sample description, no identifying data). Role; years of experience; main jurisdiction(s); how often you search case law (daily / weekly / monthly / rarely).

Section 2: Relevance judgements (blind comparison). For each query you see a short legal situation and two result lists (A and B). Rate each result and give a one-line reason. Scale: 0 = not relevant · 1 = weakly related · 2 = relevant · 3 = strongly relevant.

Query 1 (fact-pattern): "A public-school student is suspended for a silent protest during a school event; can the school punish the expression?"

List	Result cases	Relevance 0–3
A	A1 / A2 / A3	☐0 ☐1 ☐2 ☐3
B	B1 / B2 / B3	☐0 ☐1 ☐2 ☐3

Which list put the more useful cases nearer the top (A / B / about the same), and why? *Queries 2 and 3 repeat this format: a paraphrased holding, then a short out-of-domain probe ("car accident insurance claim").*

Section 3: The "no strong matches" signal. For an out-of-domain query like Query 3 the system can show: "No strongly-relevant precedent was found." Would you rather see that notice, the best-ranked cases anyway, or both? How valuable is an honest "no strong matches" signal in practice, and why?

Section 4: Closing. What would most determine whether you trusted such a tool's results? Responses are anonymised and used only for this academic project.

Appendix F: Raw / Anonymised Data Extracts

The corpus and evaluation data are public US court records with no personal data. Extracts below are verbatim from the versioned evaluation files (queries.jsonl, qrels.jsonl, ood_queries.jsonl).

Example evaluation queries (one per bucket):

ID	Bucket	Query (abridged)
q01	fact_pattern	"A man in a public phone booth … agents attached a recording device outside … without a warrant."
q03	fact_pattern	"A poor man charged with a felony asked the

ID	Bucket	Query (abridged)
		court to appoint a lawyer … the judge refused.”
q05	fact_pattern	“Black children were forced into separate public schools because of their race.”

Example graded relevance judgements (qrels for q01):

query_id	opinion_id	grade
q01	419373af…	3 (strong)
q01	236ee1f7…	1 (weak)
q01	e3d72af3…	0 (none)

Example out-of-domain probes (expected “no strong matches”): “car accident insurance settlement”, “divorce child custody relocation”, “landlord withholding security deposit”, none answerable in a SCOTUS corpus.

Appendix G: Project Logbook / Reflective Journal Extracts

Selected entries; the full record is the project’s git history, ADRs and mistakes/ log (M001–M023).

- **1 June 2026:** Phase-1 decisions locked: OpenAI text-embedding-3-small (1536-d), three-table schema, matched-snippet headline (ADR-0004); re-dimensioning was free because the chunk table was empty.
- **4 June 2026:** S1 baseline frozen (56 queries, 1,276 pairs). Hybrid RRF did not beat semantic, contradicting an assumption already in production (M007), so work re-prioritised toward reranking.
- **6 June 2026:** the AP-R reranker reached nDCG@10 0.902 (+0.085, interval excluding zero) with no per-bucket regression; the “no strong matches” signal separated in- and out-of-domain queries at AUC = 1.0.
- **9 June 2026:** the edge-function monolith was refactored under a parity gate (byte-identical responses, nDCG@10 0.898). A full ingest hit CourtListener’s undocumented 125-per-day quota (M018), stretching corpus scaling from days to weeks.
- **12 June 2026:** cases first read as “undecided” (M022) were re-diagnosed as a date-mapping gap (M023): the decision date lives on the opinion cluster, not the docket. The same day, honest result counters (relevant / shown / corpus) shipped.

Appendix H: Supervisor Meeting Records

Date	Agenda / Items discussed	Decisions and action points
[author to complete: date]	Proposal scope; aim, objectives and RQs; capstone format	[author to complete]
[author to complete: date]	Methodology and evaluation design; LLM-as-judge validity concern	[author to complete]
[author to complete: date]	Draft review; findings and discussion; limitations framing	[author to complete]

Appendix I: Assessment Criteria Mapping (for self-check)

Code	Criterion	Evidenced in Section	✓
P1	Clear aim and objectives for a complex problem or opportunity.	1.2, 1.3, 1.4	✓
P2	Significance of the project in its digital-technologies context.	1.6	✓
M1	Justify relevance, feasibility and significance from academic/industry sources.	Ch 2 (2.2–2.5), 1.6	✓
D1	Evaluate alternative approaches, with wider context.	2.5	✓
P3	Structured plan: timelines, resources, risks and ethics.	3.4–3.7	✓
P4	Implement key elements of the plan to defined outcomes.	3.9, Ch 4	✓

Code	Criterion	Evidenced in Section	✓
M2	Monitor progress with appropriate tools; respond to challenges.	3.8	✓
D2	Critically assess planning and management, with lessons learned.	3.10	✓
P5	Apply data collection and analysis to generate aligned findings.	4.1–4.3	✓
M3	Interpret appropriate data collection and analysis methods.	4.2	✓
M4	Compare patterns/trends to draw conclusions, with visualisation.	4.3, 4.4 (Figs 4.1–4.3)	✓
D3	Evaluate validity and reliability of findings, with implications.	5.3, 5.4, 5.7	✓
P6	Present outcomes in a structured report and oral presentation.	Whole report; viva separate	◑
P7	Reflect on development using a recognised model; strengths/improvements.	7.1, 7.2	✓
M5	Communicate to a professional audience, with accurate citation.	Whole report; References; Figures/Tables	✓

Code	Criterion	Evidenced in Section	✓
D4	Critically review development using evidence; propose growth strategies.	7.3–7.5	✓

Self-check: all Pass, Merit and Distinction criteria are evidenced above. The report component of P6 is complete; the oral viva is scheduled separately.