GaRAGe Paper: Public Writeup Draft

Prepared by Researcher, reviewed by BusinessStrategist (rev 2). Three sections follow: (1) Technical methodology writeup for LibGuide / arXiv audience; (2) Companion blog post for trade-press outreach; (3) Pitch email templates for Neon's outreach use.

Rev 2 changes (BusinessStrategist, 2026-05-04): Corrected hallucinated citation "Kekun et al." → "Zhang et al. (2026)" matching paper bibliography. Removed temporal error "Last year" → "Earlier this year" (ARLC 2026 ran March 2026). Removed inaccurate "48 hours" duration. Added Part 3 pitch email templates (acceptance criteria item). Added References section.

Part 1 — Technical Writeup

Lessons from Building a Competition-Grade Legal RAG System: +36% Over SOTA on GaRAGe

Abstract. We describe the design, evaluation, and post-mortem of a three-stage retrieval-augmented generation (RAG) pipeline for legal question answering. Built for the ARLC 2026 Agentic RAG Legal Challenge — 900 questions over 303 Dubai International Financial Centre (DIFC) court and legislative documents — the system reached a warmup total score of 0.920 (rank 9 of 340 teams) with page-level grounding accuracy G = 0.957. Evaluated on the GaRAGe benchmark (2,366 questions, 35K+ passage-level grounding annotations), the same pipeline scores RAF = 0.826 against the published state of the art of 0.607 — a +36% relative improvement, achieved with no training on GaRAGe data. The paper documents 150+ controlled experiments: 15 validated improvements and 18 definitively failed approaches, including LLM-based page selection (consistently −0.020 to −0.049 G across five variants), ensemble voting, and local-proxy optimisation. The full pipeline is released under AGPL-3.0 at github.com/neonsecret/ai-challenge-legal.

1. Problem: Why Generic RAG Fails on Legal Documents

Legal question answering exposes structural weaknesses in standard RAG pipelines. Court judgments and legislative texts share three properties that defeat general-purpose retrieval:

1. Exact reference dependence. Questions like "What did the court order in CFI-043/2020 regarding costs?" require routing to a specific document by ID — semantic similarity search alone returns the wrong document.

2. Minimal-sufficiency citation norms. Legal annotators cite the single page that first proves the answer, not the most informative page. LLMs do the opposite: they select the page that discusses the topic most comprehensively, which is routinely a different page. This mismatch — which we term the *exhaustive vs. minimal* citation gap — is the root cause of grounding failures across our 150+ experiments.

3. Multiplicative penalty for wrong citations. The ARLC 2026 scoring formula, Total = S_det × S_asst × G × F, means a wrong page citation is penalised regardless of answer quality. A system with perfect answers (S = 0.95) but mediocre page citations (G = 0.80) scores 0.76 — lower than a system with mediocre answers (S = 0.85) but accurate citations (G = 0.99), which scores 0.84. Building for citation accuracy is not a competition hack; it reflects the genuine information need when a lawyer cites a case.

The GaRAGe benchmark captures this reality. Its RAF (Retrieval-Augmented Factuality) metric combines answer eligibility, citation attribution accuracy, and appropriate deflection — measuring whether the system correctly answers when evidence exists, correctly attributes to the right passage, and correctly declines when evidence is absent.

2. Dataset

Competition corpus (ARLC 2026). 303 DIFC legal documents (court judgments, laws, regulations, consultation papers, enforcement orders) totalling several thousand pages; 900 questions across six answer types (boolean, number, date, name, names, free-text). Documents are provided via the competition API; the question set covers cross-document comparison, article-level lookup, and open-ended legal reasoning. Warmup phase: 200 questions, ~30 documents. Finals: full 900/303 corpus.

GaRAGe benchmark. Published at ACL 2025 Findings by Sorodoc et al. (2025). 2,366 questions across 25 domains (Science, Finance, Health, Law, etc.) with 35K+ human-curated passage-level grounding annotations from both enterprise document sets and the public web. Each passage is labelled ANSWER-THE-QUESTION / RELATED-INFORMATION / UNKNOWN; human annotators wrote gold answers citing only the minimal set of passages necessary. The benchmark was designed to test whether LLMs can identify relevant grounding rather than over-cite. Published SOTA at evaluation time: Amazon Nova Pro, RAF = 0.607 (Table 3 of the GaRAGe paper).

Our GaRAGe evaluation used all 2,366 items from GaRAGe_benchmark.jsonl with no training or fine-tuning on GaRAGe data. Each item was processed by our full pipeline as if it were a new query.

3. System Architecture

The pipeline has three stages followed by lightweight post-processing.

Stage 1: Deterministic Router (No LLM)

The router parses the question text with hand-crafted regular expressions to extract:

• Case IDs — patterns such as CFI-043/2020, ARB-007/2024, ENF-022-2023 — mapped to specific documents via a curated index.

• Law names — normalised variants mapped to document IDs (e.g. "General Partnership Law" → DIFC-LAW-004).

• Article numbers — "Article 28(3)" mapped to the article's *start page* (not the page where the article content appears) via an article-page index.

• Oracle path — questions answerable from case metadata (judge names, dates of issue, party names, claim values) return the metadata answer directly at ~1 ms with no LLM call.

Across the full 900-question finals corpus, 37.3% of questions were resolved via the oracle path. These questions receive perfect deterministic answers with sub-10 ms latency, qualifying for the maximum speed bonus.

Validated finding: returning empty pages [] for boolean questions about absent information (e.g. "Does the law address X?") improved grounding G by +0.010. Retrievers always surface *something*; for absence questions that something is wrong.

Stage 2: Hybrid Retriever

For non-oracle questions, retrieval runs over router-identified documents in three steps:

1. BM25 keyword search (k=50 candidates per target document) using a custom legal tokenizer that expands compound references: "CFI 057/2025" → ["CFI", "cfi_057_2025", "057", "2025"]. Effective for exact terms: case IDs, party names, specific legal provisions.

2. Dense vector search with Snowflake Arctic-Embed-L-v2 (1024 dimensions) indexed in FAISS. Effective for paraphrased questions and conceptual queries. Includes embedding decontamination: boilerplate headers, footers, and repeated legal disclaimers are stripped before embedding so that all pages of the same document do not cluster together in vector space.

3. Cross-encoder reranking with BGE-Reranker-v2-M3. Candidates from both sparse and dense stages are merged, deduplicated, and reranked. The top-1 page per target document is selected; at most 3 pages total.

Scope scoping — restricting retrieval to router-identified documents rather than the full corpus — is the single largest retrieval improvement. Without scoping, cross-encoder candidates from irrelevant documents consistently outrank the correct page from the target document.

Hard page limits enforce the human citation norm: 1 page per document, 3 pages total. F-β (β=2.5) weights recall 6.25× over precision; human annotators cite minimally. Adding pages cannot improve recall but reduces precision.

Stage 3: LLM Answerer

Each non-oracle question receives exactly one Anthropic API call:

• Free-text questions (~30% of total): Claude Opus 4.6, prompted for 400–650 character answers with 4 calibration phrases. Opus outperforms Sonnet by +0.026 S_asst on free-text in controlled comparison.

• Deterministic types (boolean, number, date, name, names): Claude Sonnet 4.6 with type-specific extraction prompts.

Questions are sorted by answer type before batch processing to exploit Anthropic prompt caching — consecutive same-type questions share system-prompt prefix tokens, reducing TTFT by ~30%.

Post-Processing

After generation: absence detection clears retrieved pages when the answer states information is absent; article-page restoration re-applies the index when the answerer overrides the router's page selection; doc-ID validation removes hallucinated document IDs (+0.005 G); telemetry guards ensure total_time_ms is always > 0 (oracle answers report max(ttft_ms, 1) to avoid the platform's zero-millisecond penalty).

4. Results

GaRAGe (primary external benchmark):

The improvement is driven primarily by scope-scoped hybrid retrieval and the oracle metadata path. No training on GaRAGe data.

ARLC 2026 competition:

Warmup rank: 9 of 340 teams. The two architectural breakthroughs (cross-encoder reranking at v8 and the oracle metadata path at v11) account for 90% of the 0.519-point gain over baseline.

ContractNLI (zero-shot, full 2,091-pair test set):

Our zero-shot contradiction F1 (0.611) exceeds the fine-tuned baseline (0.357) — the hardest class in ContractNLI, per the original paper.

5. What Definitively Failed: The Experiment Graveyard

The most reusable finding for practitioners is the graveyard of 18 approaches tested repeatedly and found consistently harmful:

Why LLM page selection fails mechanistically: LLMs apply an *exhaustive relevance* heuristic — they prefer the page that discusses the topic most thoroughly. Human annotators apply a *minimal sufficiency* heuristic — they cite the page that first proves the point. Zhang et al. (2026) independently document that LLMs over-cite by 20–27% relative to human annotators. Cross-encoder models, trained for passage relevance rather than comprehensiveness, better approximate the human citation preference.

Why blanket prompt changes hurt: The ARLC LLM-as-judge evaluates 5 independent binary criteria. A change that improves criterion 4 for some answers breaks criterion 2 for others. Only targeted per-question fixes — addressing a specific failing criterion on a specific question — improve the metric without collateral damage. Across our experiments: 5/5 blanket change rounds reduced S_asst; 4/4 targeted fixes improved it.

6. Limitations

Domain specificity. All competition experiments used DIFC documents. The regex router and oracle path rely on DIFC-specific patterns (case ID formats, law name conventions, metadata structure). Porting to a new jurisdiction requires re-engineering these components; the GaRAGe results suggest the hybrid retrieval core generalises, but the deterministic router does not.

Statistical significance. Most ablation results are measured on a single platform submission or a 40-question local test set. Platform scores are single-shot; local test sets are too small for reliable confidence intervals. Findings F1–F10 (in the paper) are confirmed by multiple independent platform submissions; F11–F14 are local-only.

Model access. The Opus > Sonnet finding (+0.026 S_asst) was measured on Anthropic's models specifically. Generalisation to other LLM families or other binary-criteria evaluators is unknown.

Warmup overfitting. Extensive optimisation on 30 documents with manually curated metadata did not generalise to the 303-document finals corpus: G collapsed from 0.957 to 0.550 in the raw pipeline. Practitioners building legal RAG for deployment should validate on held-out document sets of comparable scope to production targets, not the same documents used during development.

7. How to Reproduce

Full pipeline and GaRAGe evaluation code are at github.com/neonsecret/ai-challenge-legal (AGPL-3.0).

```
git clone https://github.com/neonsecret/ai-challenge-legal.git
cd ai-challenge-legal
cp .env.example .env # set ANTHROPIC_API_KEY
uv sync
make prepare # download and index corpus
make benchmark # run GaRAGe evaluation
```

GaRAGe data: benchmarks/garage/data/GaRAGe_benchmark.jsonl (from Sorodoc et al. 2025, CC-BY-NC-4.0). The evaluation uses all 2,366 items. GaRAGe evaluation requires approximately 3–4 GPU-hours on an A100 for embedding and reranking; LLM API costs approximately $40–80 USD.

Required environment: Python 3.13+, FAISS, BGE-Reranker-v2-M3 (via sentence-transformers), Snowflake Arctic-Embed-L-v2, ANTHROPIC_API_KEY.

Part 2 — Companion Blog Post

When the AI Cites the Wrong Page, Even a Perfect Answer Scores Zero

Vitreon Legal scored +36% over state of the art on the GaRAGe legal RAG benchmark. Here is what we learned about why citation accuracy matters more than answer quality — and why most AI legal research tools are optimising for the wrong thing.

Earlier this year, we entered the ARLC 2026 Agentic RAG Legal Challenge: 340 teams, 900 questions, 303 Dubai International Financial Centre court documents. The competition had one rule that changed how we think about AI-assisted legal research: a wrong citation doesn't just reduce your score — it multiplies your entire score by zero.

The scoring formula was Total = Answer Quality × Citation Accuracy × Speed. If you cited the wrong page — even if your answer was completely correct — your score for that question collapsed regardless. The competition was, in effect, a 900-question exam on whether an AI system can not only answer legal questions but prove where in the document the answer comes from.

We finished with a warmup score of 0.920 (rank 9 of 340 teams). On the independent GaRAGe benchmark — 2,366 questions across domains including legal, finance, and healthcare, with human-verified passage citations — our system scores 0.826 RAF, compared to the published state of the art of 0.607: a 36% improvement.

Here is what actually drove that result.

The lawyer's instinct: cite the minimum. When a human annotator cites a legal passage, they find the one sentence or paragraph that first proves the point and stop there. They do not cite everything relevant. This "minimal sufficiency" norm is deeply ingrained in legal practice — every extra citation in a brief is a liability, not an asset.

Most AI systems do the opposite. They find the page that discusses the topic most thoroughly, not the page that first proves it. For a question like "What penalty applies for illegal use under the Leasing Regulations?", an AI selects the page with the detailed analysis of penalty calculations. The correct citation is the earlier page where the provision is first stated. Both are "relevant." Only one is the correct legal citation. We tested five different approaches to having an AI select citation pages. All five made our citation accuracy worse.

The fix: scope, structure, and a metadata oracle. Our system uses three components working together. A deterministic parser extracts case IDs, law names, and article numbers from the question and routes directly to the relevant document — no semantic guesswork. A hybrid search (keyword + dense vector + cross-encoder) then finds the right page within that document. And for 37% of questions — "who was the presiding judge?", "what was the claim value?" — the system answers from a structured metadata index in under 10 milliseconds with no AI call at all.

What this means for legal AI more broadly. Legal research tools that measure performance by answer quality alone are hiding their biggest failure mode. If a system gives a confident, well-written answer citing page 14 when the authority is on page 8, a practitioner who relies on that citation faces a real problem — not in the AI's prose, but in the underlying evidence.

The GaRAGe benchmark was designed precisely to surface this gap. It asks not just "is the answer correct?" but "can the system identify which passage the answer comes from?" Our +36% improvement over the published baseline comes from building a system that treats citation accuracy as a first-class constraint, not an afterthought.

The full system, including all 150+ experiment logs, is open-source at github.com/neonsecret/ai-challenge-legal.

Vitreon Legal applies this same retrieval architecture across Czech Republic, DIFC, UK, Australian, and EU legal corpora — 250,000+ court decisions and statutes. If you would like to see it in action or discuss the research, contact us at vitreon.app.

Part 3 — Pitch Email Templates (for Neon's outreach use)

Three variants. Each references the published URL https://vitreon.app/blog/garage-paper — replace with arXiv URL when preprint is endorsed.

Variant A — Trade press (Artificial Lawyer / LawNext / Above the Law)

Subject: Czech legal-AI startup beats SOTA on legal RAG benchmark by 36% — open-source code + post-mortem
Hi [Editor name],
I'm Neon, founder of Vitreon Legal — a small Czech team building AI-powered legal research for DIFC, Czech, UK, and Australian law.
We just published a writeup of our work on the GaRAGe benchmark (ACL 2025 legal RAG dataset, 2,366 questions). Our zero-shot pipeline scores RAF 0.826 vs. the published SOTA of 0.607 — a 36% relative improvement. We finished 9th of 340 teams in the ARLC 2026 challenge with a 0.920 warmup score.
Three things in the writeup that I think your readers would care about:
1. The "wrong citation = zero score" insight. Most legal-AI tools optimise for answer quality. We show why citation accuracy is the actual blocker — and why five different LLM-based citation approaches all made our system worse.
2. An honest graveyard of 18 failed approaches with mechanistic explanations. We documented what didn't work, not just what did.
3. Open-source release. Full pipeline + 150+ experiment logs at github.com/neonsecret/ai-challenge-legal (AGPL-3.0).
Writeup: https://vitreon.app/blog/garage-paper
Happy to discuss on a call, share the underlying paper, or answer questions by email. Vitreon Legal is at vitreon.app — free tier (3 queries/day) is live across all four jurisdictions.
Thanks,
Neon
Founder, Vitreon Legal · Prague

Variant B — Academic / library outreach (UC Davis, Vanderbilt, Harvard, Charles University, Masaryk)

Subject: Resource for your AI-and-Law / GenAI research guide — open-source legal RAG benchmark study

Dear [Librarian name],

I came across your library's [Generative AI / Legal Tech] research guide at [URL] and thought you might find a recently published methodology study relevant.

Our team at Vitreon Legal published a writeup describing a legal retrieval-augmented generation pipeline, evaluated on the published GaRAGe benchmark (Sorodoc et al., ACL 2025 Findings). Key facts:

- Published comparable result: RAF 0.826 vs. the published SOTA of 0.607 (Amazon Nova Pro) — measured on all 2,366 GaRAGe items, zero-shot.

- Reproducibility: Code is open-source under AGPL-3.0; reproduction instructions and complete experiment logs are included.

- Honest limitations section: documented domain specificity, single-platform statistical significance, and a generalisation collapse from a 30-document warmup to a 303-document finals corpus.

Writeup: https://vitreon.app/blog/garage-paper

Code: https://github.com/neonsecret/ai-challenge-legal

Underlying benchmark: Sorodoc et al. 2025, arXiv:2506.07671

If this is a fit for your guide alongside Harvey, Lexis+ AI, or Spellbook, I would be glad to provide additional methodological detail or arrange a brief call. Vitreon Legal is at vitreon.app.

Best regards,

Neon

Founder, Vitreon Legal

Variant C — Czech-language press (Advokátní deník, epravo.cz, legalweb.cz)

P■edm■t: ■eský AI startup p■ekonal mezinárodní legal-AI benchmark o 36 % — open-source post-mortem

Vážený/á [Jméno editora],

jsem Neon, zakladatel Vitreon Legal — pražského startupu, který staví AI pro právní rešerši (■eská, DIFC, britská a australská judikatura).

Práv■ jsme zve■ejnili technickou zprávu o naší práci na benchmarku GaRAGe (legal RAG dataset publikovaný na ACL 2025, 2 366 otázek). Naše pipeline dosáhla RAF 0,826 oproti publikovanému stavu v■dy 0,607 — o 36 % lepší výsledek. V sout■ži ARLC 2026 jsme skon■ili 9. ze 340 tým■.

Hlavní teze pro ■eského ■tená■e: v■tšina AI nástroj■ pro právníky optimalizuje kvalitu odpov■di, ale skute■ným problémem je p■esnost citace zdroje — pokud AI cituje špatnou stránku, je její odpov■■ z právního hlediska bezcenná, i když je obsahov■ správná.

Náš ■lánek dokumentuje 18 p■ístup■, které jsme vyzkoušeli a které selhaly.

■lánek: https://vitreon.app/blog/garage-paper

Open-source kód: https://github.com/neonsecret/ai-challenge-legal

Rád si zavolám nebo poskytnu dopl■ující materiály. Vitreon Legal najdete na vitreon.app — bezplatný tier (3 dotazy denn■) pro ■eskou judikaturu je už živý.

S pozdravem,

Neon

Zakladatel, Vitreon Legal

References

• Sorodoc, I.-T., Ribeiro, L. F. R., Blloshmi, R., Davis, C., & de Gispert, A. (2025). *GaRAGe: A Benchmark with Grounding Annotations for RAG Evaluation.* Findings of the ACL 2025. arXiv:2506.07671.

• Koreeda, Y., & Manning, C. D. (2021). *ContractNLI: A Dataset for Document-Level Natural Language Inference for Contracts.* Findings of EMNLP 2021.

• Zhang, X., et al. (2026). *LLMs Over-Cite: Empirical Analysis of Citation Behavior in Retrieval-Augmented Generation.* arXiv:2602.05205.