We put Google's search on 480,000 German court decisions and measured it

Bork Morfaw · 2026-07-27

research.nulegal.eu/findings/google-search-vs-our-retrieval

TL;DR

- On real user queries, our engine answers 91% in the top-5; Vertex AI Search over the same corpus answers 77%; Google's web index manages 32%.
- On 2,941 held-out case-law questions, our retrieval finds the correct decision 2.2× as often as Vertex AI Search (R@10 0.61 vs 0.28).
- On 116 landmark research questions Vertex edges us out at rank 10 (0.55 vs 0.52), and the two engines find different decisions: combined they'd reach 72%. That fusion is now on our roadmap.

1. Why we ran this

Our search over gesetze.nulegal.eu (every German federal statute plus 480,000+ court decisions) is built in-house: hybrid lexical + dense retrieval with a fine-tuned German-legal query encoder, citation-graph priors, and an LLM rerank path for natural-language questions. The obvious question any team should ask itself: **would off-the-shelf Google do just as well?**

We tested both ways Google can sit on top of a corpus like ours:

1. **Google Programmable Search (CSE)** restricted to our public site: Google's web index of the corpus.
2. **Vertex AI Search** with our full corpus (584,465 documents: 100,001 statute sections with body text (§/Art./Anlage provisions rather than heading-only rows) + 484,464 decisions at benchmark time) pushed into a datastore: Google's retrieval engine, with no index-coverage excuse.

2. Setup

Three evaluation sets, none synthetic:

- **Organic probe**: 119 real user queries from our search logs, stratified by the query styles people actually type (citations, keywords, lay questions, docket numbers, typos). Judged by Gemini 2.5 Pro at temperature 0, with a second pass at temperature 0.7 as a cross-check (agreement on the headline metric: 0.90–0.92). The judge is Google's own frontier model, a deliberate choice: if the judging leans anywhere, it leans toward the Google baselines, not us. Metric: answered@5. Would the searcher's information need be satisfied by the top-5 results?
- **case_eval**: 2,941 held-out questions, each with one known-correct decision from 2026. Judge-free: the gold decision is either in the top-k or it isn't.
- **landmark-v2**: 116 verified research questions, each targeting one landmark decision (our permanent holdout set).

3. Results

eval	metric	Google CSE	Vertex AI Search	ours
organic probe (119)	answered@5	0.319	0.773	**0.908**
case_eval (2,941)	recall@10	n/a[1]	0.275	**0.610**
landmark-v2 (116)	recall@10	0.069[2]	**0.552**	0.517
landmark-v2	recall@20	0.069[2]	0.621	**0.647**

[1] CSE cannot be scored on case_eval: the gold decisions are largely absent from Google's index of our young site, so a recall number would measure crawl coverage, not retrieval. [2] CSE returns at most 10 results per query, so recall@20 cannot exceed recall@10.

The CSE number is an index-coverage story, not a retrieval story. As of July 2026 the site is weeks old; Google's crawl covers a small fraction of it. 74 of the 116 landmark questions returned zero results. We'll re-run this quarterly as the index grows. As of today, "just put a Google search box on it" answers one query in three.

Vertex AI Search is a serious baseline, and it still loses on what users actually type. With the full corpus indexed, it answers 77% of real queries against our 91%, and it trails badly on precise case-law lookup (0.275 vs 0.610 recall@10). Domain-tuned retrieval (the citation graph, the fine-tuned encoder, German legal query handling) earns its keep exactly where precision matters.

The interesting result is the one we lost. On long, doctrine-heavy research questions, Vertex's semantic retrieval finds landmark decisions our stack misses at rank 10. The headline gap (0.552 vs 0.517) is four questions out of 116, well within sampling noise at this size; what makes the result real is the diff, not the delta. The diff is what makes it actionable: of 116 questions, both engines find 40, only ours finds 20, only Vertex finds 24, neither finds 32. The engines are complementary: a fused candidate pool would lift recall@10 from 52% to a 72% ceiling. That fusion experiment is next on our roadmap.

4. Method notes

- The case_eval pair is same questions, same gold decisions, same metric; the measurement paths differ. Vertex was run against its live API; ours was measured offline against the production dense retriever (the same encoder and index the live system serves), without the serving layer on top. The landmark rows are both live-API runs.
- Judged runs follow our standing rules: strong judges only, two independent passes, structured output, no cherry-picking. The full methodology is on the methodology page.
- The Vertex datastore held a same-day snapshot of exactly the documents our engine searches, so corpus coverage is identical by construction.
- Two judge calls on the organic probe were arguably generous to Google (accepted answers on shaky rationales); we count them anyway. The gap doesn't depend on them.
- Costs, for anyone repeating this: the entire exercise (datastore, 10.6 GiB import, ~6,200 queries across all runs) cost under €30 in API fees.

5. Reproduce it

The harness (runner scripts, judge prompts, metric code, and the evaluation sets) is published in our research repository under research-friendly licenses (code: PolyForm Noncommercial; data: CC BY-NC 4.0; commercial licensing on request). German court decisions themselves are in the public domain (§ 5 UrhG).