Privileged

Accuracy Study — Local AI Legal Document Review

August 2026 · getprivileged.ai/accuracy · App version 1.0.1

Study date: August 2026 · App version: 1.0.1 · Every number on this page comes from a logged run of the shipping application. Nothing here is estimated.

Headline results — current default model (Gemma 4 E4B, hardened pipeline)

Benchmark: 150 real commercial contracts (CUAD, expert-annotated) · 10 questions each · run in the shipping app · 1,310 answers scored against the expert answer key.

Metric	Result
Coverage — clauses that exist, found	**64.5%**
Verified-answer accuracy	**97.2%**
False “verified” answers	**7 of 1,310 (0.53%)** — every one hand-reviewed and described below
Correct “not present” on absent clauses	**94.3%**
Adversarial near-miss traps correctly refused	**10 / 10**
Everyday documents (leases, standard agreements)	**0 fabrications; every verified answer correct** in our testing

How to read these numbers: Privileged is a triage tool. A verified answer carries a quote and page number you can check in seconds, and is right ~97% of the time. An unverified or empty cell means *check this yourself* — it never means “this doesn’t exist.” The tool does the first pass with receipts; you do the judgment.

Methodology

Dataset. CUAD v1 — the Contract Understanding Atticus Dataset (Hendrycks et al., NeurIPS 2021): 510 real commercial contracts annotated by attorneys across 41 clause categories. License CC BY 4.0. Archive verified against the published Zenodo checksum (MD5 `c38f490a984420b8a62600db401fafd5`).

Selection. 150 contracts drawn by seeded random selection, stratified by document length (three equal buckets). The seed and full document manifest are published — anyone can reproduce the exact sample. We did not choose the documents.

Questions. Ten questions, phrased the way an attorney would type them, frozen before the run:

1. What law governs this agreement?
2. Can either party terminate this agreement for convenience?
3. Is there a cap on either party’s liability?
4. Does this agreement restrict either party from competing?
5. What happens to this agreement if a party is acquired?
6. Can this agreement be assigned without consent?
7. Does the agreement renew automatically?
8. Does either party have audit rights?
9. What is the expiration date of this agreement?
10. Is there an exclusivity provision?

Execution. Every run was performed in the packaged shipping application — the same build users download — via UI automation. No harness-only shortcuts, no study-specific tuning. Documents were processed as size-stratified reviews, matching how the app batches real work.

Scoring. Model answers were scored against CUAD’s expert annotations: normalized answers for categorical questions, annotated clause spans for text questions, plus an independent, deterministic check that every cited quote actually appears on the cited page of the source document.

Hand review. Every cell the scorer flagged as a false “verified” answer was individually reviewed by a human against the source document and classified: scorer artifact (the model was right, the string-matching was pedantic), true error (by type), or ground-truth dispute (we believe the answer key itself is wrong — with document evidence quoted). We publish all three resulting readings of the

false-verified rate rather than only the most favorable one.

Full results — default model

Run accounting. Of 150 selected contracts: 131 completed fully · 13 excluded up front as exceeding the current document-size limit (the app flags these before running rather than silently truncating; corpus-wide, 13.2% of CUAD exceeds the current limit) · 6 failed with engine errors on very long documents, isolated to error cells (error cells cannot produce false answers).

False-verified accounting — all three readings:

Reading	Count	Rate (of 1,310 scored)
Strict (every scorer flag counted)	17	1.30%
Middle (scorer artifacts excluded, disputes counted against us)	11	0.84%
Signed (hand-reviewed)	**7**	**0.53%**

Of the 17 flagged cells: **6 were scorer artifacts** — in five, the model quoted the *identical clause* the experts annotated and answered correctly, differing only in label ("Swiss law" vs "Switzerland"); **7 were true errors** (detailed below); **4 were ground-truth disputes** (detailed below).

The 7 true errors, individually:

1. A signature-block date cited as the agreement's expiration date (the real answer required the term clause).
2. Generic successors-and-assigns boilerplate cited as a change-of-control provision.
3. A clause about acquiring third-party assets cited as if it covered a party being acquired.
4. A limitation-of-liability clause phrased in the negative, inverted — the model read "no liability" as "no cap." *(The one negation-class error in the run.)*
5. An exclusivity answer that extracted a qualifying sentence while dropping the operative grant one sentence earlier.
6. A multi-jurisdiction agreement where the model returned the pointer clause rather than the specific governing laws.
7. A merger-survival clause quoted for a change-of-control answer where the annotation pointed to a different provision. *(Borderline — a real, on-point clause; counted as an error under our conservative rule.)*

Every one of these produced a checkable citation — a lawyer clicking through would land on the quoted text and could evaluate it directly.

The 4 ground-truth disputes — where we believe the answer key is wrong:

The strongest: CUAD marks one agency agreement's term as *perpetual, no expiration*. The document states, verbatim: **"Agency Agreement Expiration Date and Last Delivery Date: October 31, 2006."** The model reported October 31, 2006. We scored this as a dispute, not a win — but we quote the document and let you judge. The other three are category-boundary judgments (e.g., a restriction on competing endorsements that CUAD files under exclusivity rather than non-compete). Disputed cells were **excluded from our favor in the signed reading only where the dispute was sustained; all are disclosed here.**

Abstention accounting. The verification stack demoted 107 answers to "unverified" during the run: 47 were genuine catches (the answer or its support didn't hold up) and 60 were over-caution — correct answers demoted anyway. Over-caution costs you one click to check a right answer; it is the failure direction we deliberately chose. Most over-caution concentrated on two yes/no questions.

Adversarial testing. Beyond CUAD: a 19-shape family of negation traps (clauses that *deny* what a skim suggests they grant) — zero false answers; 10 near-miss absence probes (a related-but-wrong clause present: termination-for-*cause* when asked about *convenience*, forum selection when asked about governing law, and eight more) — 10/10 correctly refused, while the neighboring clauses still extracted correctly, confirming discrimination rather than blanket caution; and 4 real leases and an amendment run through the app — zero fabrications, every verified answer matching ground truth, including correctly refusing to call 2×5-year renewal *options* "automatic renewal."

The model story — why the pipeline matters more than the model

Our original default (Qwen3 4B) scored near-perfect on clean synthetic test documents for two years — then found **2%** of what real CUAD contracts contain. Dense real-world legal prose triggered a failure our synthetic suite never exercised. What held: in 109 failed answers, it fabricated **zero** — the verification requirement that every answer carry a real quote from the actual page meant it failed silent, not false.

The replacement (Gemma 4 E4B) failed in the opposite direction in initial qualification: confidently citing real sentences that didn't support its answers. Real quotes, wrong meaning — a failure class quote-checking alone cannot catch.

The response was a hardened pipeline: model-specific handling of how each model is fed and constrained, plus additional verification

layers that check not only that a quote exists but that it supports the claimed answer. The specific mechanisms are proprietary. The results are the tables above — and the qualification gate that produced them is permanent: **no model ships as a Privileged tier without passing this full benchmark, and every future model's results will be published on this page.**

We also ran the experiments that *didn't* work, and report that too: one candidate verification design missed its acceptance targets and was shelved; one prompt-level intervention that fixed one model measurably broke the other. Negative results are part of the record.

Hardware (measured, not estimated)

	Default (Gemma 4 E4B)	Legacy (Qwen3 4B)	Heavy (Qwen3 14B)
Download	~8.9 GB	~2.5 GB	~9.3 GB
RAM	8 GB minimum · **16 GB recommended for long documents** (measured peak on the longest bucket: 7.9 GB)	8 GB	32 GB recommended
Typical speed	~50s per document (10 questions, full verification) · ~90s on very long documents	—	—

Documents past roughly 25 dense pages may require 16 GB or be flagged as exceeding current limits. The app tells you before running. It never silently truncates.

Limitations — read this section

- **Coverage is ~two-thirds on dense commercial contracts.** An empty cell is an instruction to look, never a conclusion that nothing is there.
- **The Heavy tier (Qwen3 14B) has not yet completed this full 150-document gate.** Its published numbers are pilot-scale until it does. *(Scheduled — see changelog.)*
- **Scanned PDFs are not yet supported.** Text-layer PDFs only.
- **Registers we have not yet tested:** immigration filings and court documents. Our benchmark covers commercial contracts and real leases. We will say so here when that changes.
- **Known residual error classes exist.** One sentence-shape still produces occasional errors under adversarial conditions (3 cells in 195 on our internal adversarial suite); our synthetic fixtures pin regressions but do not prove the messier wild variants are fixed. There are likely error shapes we have not found yet.
- **One confirmed coverage inconsistency:** an assignment clause found in one lease was missed in another with similar structure. Coverage-side (an abstention, not a fabrication), logged.
- **We did not benchmark against cloud models.** This study measures whether the tool does the work, not how it ranks.
- **Very long documents** are currently excluded above a size limit (13.2% of the CUAD corpus). Removing this limit is our next engineering milestone, and when it ships we will re-run this exact benchmark and publish before/after here.

Changelog

Date	Entry
Aug 2026	Initial study published. Gemma 4 E4B qualified and shipped as default.
Planned	Long-document support → full benchmark re-run, before/after published
Planned	Heavy tier (Qwen3 14B) through the complete 150-document gate
Planned	Next model qualification (each new model runs this exact gauntlet before shipping)
Planned	Additional document registers (immigration, hand-verified answer keys)

Reproduce it

1. Download CUAD v1 (free, public, linked above) and verify the checksum.

2. Download Privileged (getprivileged.ai) — the shipping app is the benchmark environment; there is no special build.
3. Use our published seed and manifest to assemble the same 150 documents, or pick your own.
4. Run the ten questions above in a review. Compare against CUAD's annotations.
5. Grade us yourself. If you find something we got wrong, email akash@getprivileged.ai — disputes get investigated and published here.