

Legal AI tests and studies, read Oct 24, 2026

Public benchmarks of legal AI, summarised from their primary pages and read on Oct 24, 2026: who measured, when, which tools, what is measured and where the method stops. The figures are the publishers', not ours.

Online version, data and updates: <https://aiforlawyers.app/report/legal-ai-benchmarks>

CITE THIS

aiforlawyers.app (AI Vortex), Legal AI benchmarks, read Oct 24, 2026,
<https://aiforlawyers.app/report/legal-ai-benchmarks>, licensed CC BY 4.0.

Each figure belongs to its publisher; cite the primary source linked in its row.

The benchmarks, side by side

not our score

Six public benchmarks, read on Oct 24, 2026. Open a row for tools, metric, result as published, limit and the primary link.

Benchmark	Who measured it	Date	What it tests
Vals Legal AI Report (VLAIR), first edition	Vals AI, with a lawyer control group (the Lawyer Baseline)	Feb 27, 2025	legal products, 7 tasks
Vals Legal AI Report: legal research (VLAIR legal research)	Vals AI, with a lawyer control group and six law firms that wrote the questions	Oct 14, 2025	legal products, research
Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools (Stanford)	Magesh, Surani, Dahl, Suzgun, Manning and Ho (arXiv preprint)	May 30, 2024	legal research products, hallucination
LegalBench (paper and open task set)	Neel Guha and 39 other authors (arXiv preprint), built with legal professionals	Aug 20, 2023	language models, legal reasoning
LegalBench on Vals AI (live scores)	Vals AI (runs the LegalBench task set on current models)	Oct 13, 2026	language models, live scores
Legal Research Bench on Vals AI (live scores)	Vals AI (research paper linked from the page)	Oct 22, 2026	language models in an agent harness, US legal research

Each one, with its limit

Read on Oct 24, 2026 from the primary page linked in each card.

LEGAL PRODUCTS, 7 TASKS

Vals Legal AI Report (VLAIR), first edition

Measured by Vals AI, with a lawyer control group (the Lawyer Baseline) · dated Feb 27, 2025 (report page, "Updated: 02/27/2025") · read Oct 24, 2026

As published. The report states that Harvey opted into six of the seven tasks, received the top score among the participating AI tools on five and placed second on one. CoCounsel scored from 73.2% to 89.6% on its four tasks. The Lawyer Baseline outperformed the AI tools on two tasks and matched the best tool on one; one of those was redlining (79.7%).

Tools. CoCounsel (Thomson Reuters), Vincent AI (vLex), Harvey Assistant (Harvey), Oliver (Vecflow). Lexis+AI (LexisNexis) withdrew.

Metric. Accuracy or performance score against predefined criteria on seven tasks (data extraction, document Q&A, summarization, redlining, transcript analysis, chronology generation, EDGAR research).

Limit of the method. Vendors could withdraw from any task before publication, so the tools were not all measured on the same tasks (Harvey withdrew from EDGAR research, vLex from chronology generation, LexisNexis from every skill except legal research). Scoring is automated (LLM as judge). Products as of the 2025 study.

Primary source. <https://www.vals.ai/industry-reports/vlair> (read Oct 24, 2026)

LEGAL PRODUCTS, RESEARCH

Vals Legal AI Report: legal research (VLAIR legal research)

Measured by Vals AI, with a lawyer control group and six law firms that wrote the questions · dated Oct 14, 2025 (report page, "Updated: 10/14/2025") · read Oct 24, 2026

As published. Average weighted scores were 74% to 78% for the four AI products and 69% for the Lawyer Baseline. The page reports 78% to 81% accuracy for the legal AI products and 80% for the generalist product. It says Counsel Stack received the highest score across all criteria.

Tools. Alexi, Counsel Stack, Midpage and ChatGPT (OpenAI), plus the Lawyer Baseline.

Metric. 200 U.S. legal research questions scored on three weighted criteria: accuracy (50%), authoritativeness (40%) and appropriateness (10%).

Limit of the method. Only vendors that agreed to take part were evaluated, four in all. The lawyers wrote shorter answers by default, which the report says may have lowered their scores. Reference answers came from the contributing firms, and evaluators could use discretion where a reference answer was incomplete.

Primary source. <https://www.vals.ai/industry-reports/vlair-10-14-25> (read Oct 24, 2026)

Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools (Stanford)

Measured by Magesh, Surani, Dahl, Suzgun, Manning and Ho (arXiv preprint) · dated May 30, 2024 (arXiv v1, submitted 30 May 2024) · read Oct 24, 2026

As published. The authors report that the Lexis+ AI, Westlaw AI-Assisted Research and Ask Practical Law AI tools each hallucinate between 17% and 33% of the time, and that hallucinations are reduced relative to general-purpose chatbots (GPT-4).

Tools. Lexis+ AI (LexisNexis), Westlaw AI-Assisted Research and Ask Practical Law AI (Thomson Reuters), with GPT-4 as the general-purpose comparison.

Metric. Share of responses that hallucinate, from a preregistered set of legal research queries and a typology that separates hallucinated from accurate responses.

Limit of the method. The abstract gives one range for the three tools, not a figure per tool. A May 2024 snapshot of products that have changed since. A hallucination rate depends on the question set and on each product being used for what it is designed for.

Primary source. <https://arxiv.org/abs/2405.20362> (read Oct 24, 2026)

LegalBench (paper and open task set)

Measured by Neel Guha and 39 other authors (arXiv preprint), built with legal professionals · dated Aug 20, 2023 (arXiv v1, submitted 20 Aug 2023) · read Oct 24, 2026

As published. The paper presents 162 tasks in six types of legal reasoning and an empirical evaluation of 20 open-source and commercial language models.

Tools. 20 open-source and commercial language models in the paper; no legal software products.

Metric. Performance on 162 tasks covering six types of legal reasoning, written or hand-crafted by legal professionals.

Limit of the method. It measures language models, not legal products. The 20 models are those of 2023. It is a task set and a method; current scores live on leaderboards that run it, such as the Vals page in the next row.

Primary source. <https://arxiv.org/abs/2308.11462> (read Oct 24, 2026)

LegalBench on Vals AI (live scores)

Measured by Vals AI (runs the LegalBench task set on current models) · dated Oct 13, 2026 (leaderboard page, "Updated 10/13/2026") · read Oct 24, 2026

As published. As of October 13, 2026 the Vals page lists Claude Fable 5 first with 88.56%, followed by Claude Fable 5.1 (88.51%) and Gemini 4 Argon (88.30%).

Tools. Frontier and open models from many labs; the page shows 23 latest, top and frontier models in its chart.

Metric. Accuracy on open-source legal reasoning tasks.

Limit of the method. The first three scores are 0.26 points apart and the page text we read shows no interval, so the order among them is not a clear difference. Scores are for raw models, not legal products. The page changes with every new model, so read the date.

Primary source. <https://www.vals.ai/benchmarks/legalbench> (read Oct 24, 2026)

Legal Research Bench on Vals AI (live scores)

Measured by Vals AI (research paper linked from the page) · dated Oct 22, 2026 (leaderboard page, "Updated 10/22/2026") · read Oct 24, 2026

As published. As of October 22, 2026 the Vals page reports Muse Spark 1.3 Max, Claude Opus 5 and Claude Fable 5.1 tied for first at 55.29%, with Gemini 4 Argon next at 54.81%.

Tools. Frontier and open models from many labs, run as agents with case-law search, web search and document retrieval.

Metric. All-pass accuracy: a question counts only if the answer satisfies every item of a lawyer-written rubric (0% otherwise), graded by an LLM judge. A partial-credit score is also reported.

Limit of the method. Models in an agent harness, not legal products such as Harvey or CoCounsel. The test set is private, so outsiders cannot rerun it. Grading is by an LLM judge. All-pass is a strict bar, so scores look low by design.

Primary source. https://www.vals.ai/benchmarks/legal_research (read Oct 24, 2026)

How to read a benchmark

- **Products and models are different tests.** Some rows test legal products as a vendor ships them; others test raw language models. A model score says little about a product built on it, and the reverse.
- **Who took part shapes the result.** In the Vals reports vendors chose whether to take part and, in the first report, which tasks; a product that is absent tells you nothing about how it would have done.
- **The grader matters.** Several rows are graded by a language model acting as judge. Read each row's limit before you quote its number.
- **Dates matter.** Products and models change within months. A figure is only as current as the date shown next to it, and the live-score rows change with each new model.
- **We copy and attribute.** Nothing here is added to, averaged with or ordered against another row, and we do not pick a tool. Where a publisher lists an order, we say so and attribute it to them.

Our record is somewhere else

What lawyers say about these tools, and our own counts, are separate. See Arena, Frontier models for legal work, Legal research and Drafting. None of those figures is on this page, and nothing on this page is part of them.

Every figure belongs to its publisher; the primary link and the read date are in each row. The summaries are free to reuse under CC BY 4.0. Credit “aiforlawyers.app (AI Vortex)” with the date and a link to <https://aiforlawyers.app/report/legal-ai-benchmarks>.

Report generated from the pages read on Oct 24, 2026.