

Industry Reports

Comprehensive legal technology platform providing advanced research, document analysis, and case management solutions for legal professionals

Application Reports

Industry Reports

Legal Research Report



VLAIR - Legal Research

vals.ai

This study extends the initial Vals Legal AI Report (VLAIR) and evaluates how four AI products perform on legal research tasks, studying their performance with respect to a lawyer control group (the Lawyer Baseline). The evaluated products were Alexi, Counsel Stack, Midpage and ChatGPT.

Our study assessed the participants' ability to respond to 200 U.S. legal research questions, distributed across a range of question types typically encountered in private practice. We established a scoring rubric taking into account prior studies and industry feedback, ultimately evaluating the responses across three weighted criteria:

- **Accuracy:** whether the response is substantively correct with no incorrect elements (50% of average weighted score)
- **Authoritativeness:** whether the response is supported by citations to relevant and valid primary and/or secondary law sources (40% of average weighted score)
- **Appropriateness:** whether the response is easy to understand and could be immediately shareable with colleagues or clients (10% of average weighted score)

The results as well as details about the study methodology and limitations are set out below.

Executive summary

The performance across participants was broadly consistent when looking at both the final average weighted scores and the individual criteria scores.

On an average weighted score basis, all AI products scored within 4 points of each other (74%-78%) and within 9 points of the Lawyer Baseline (69%).

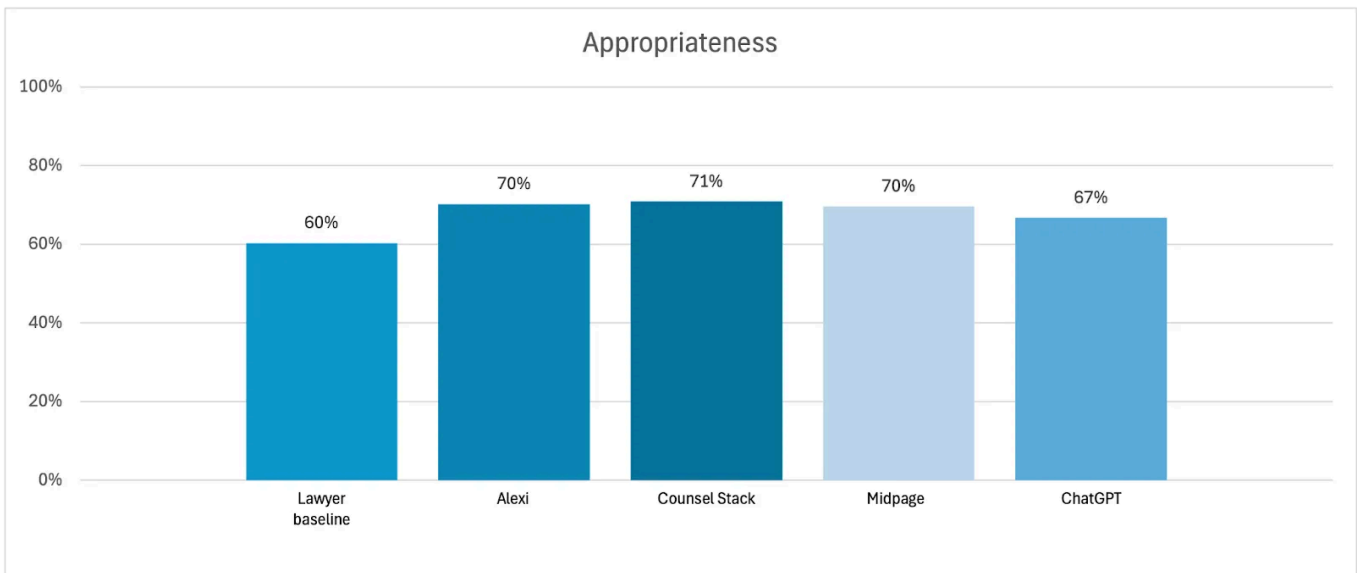
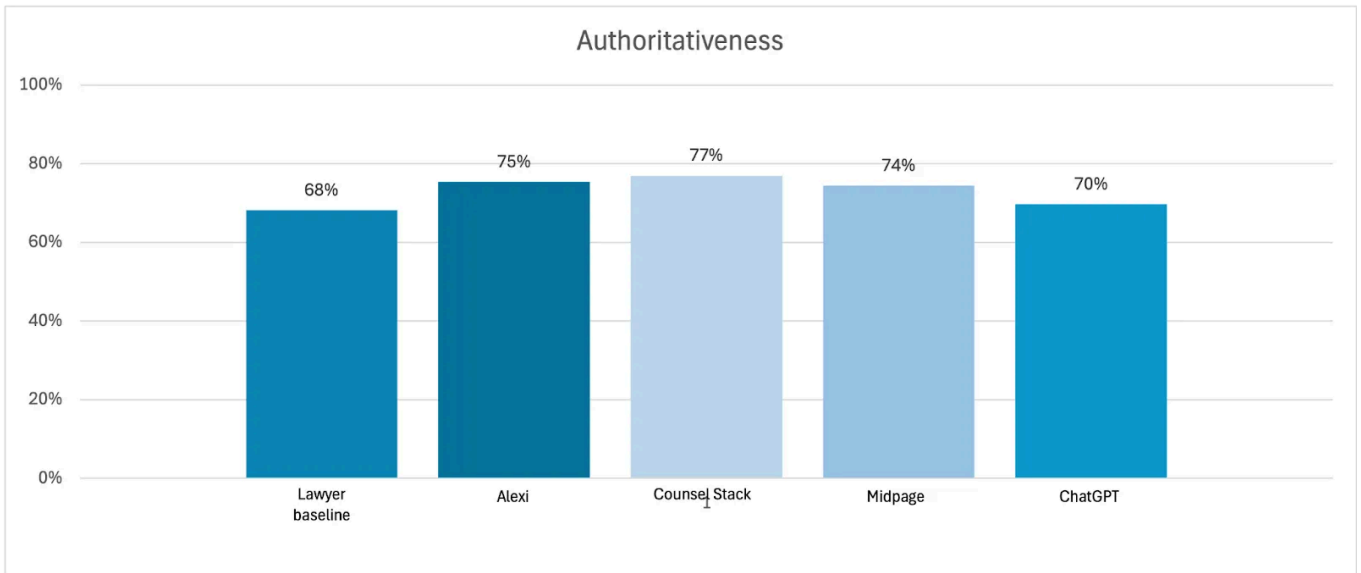
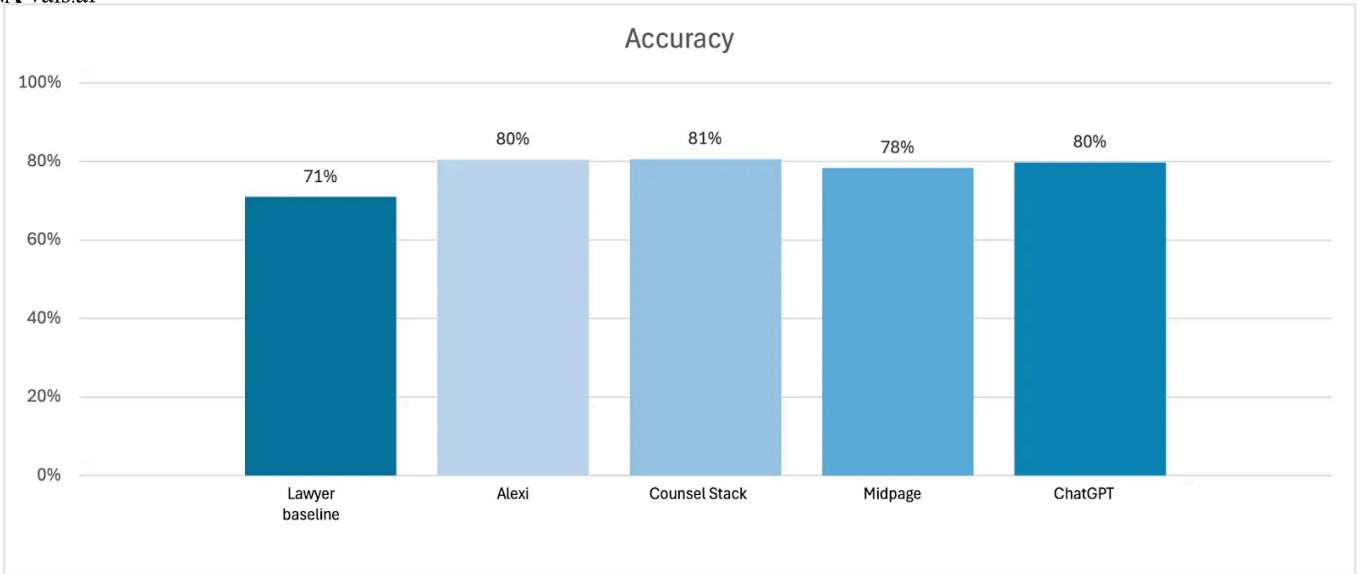
The legal AI products performed better overall than the generalist AI product, which in turn performed better overall than the Lawyer Baseline.

All of the AI products outperformed the Lawyer Baseline across the three scoring criteria.

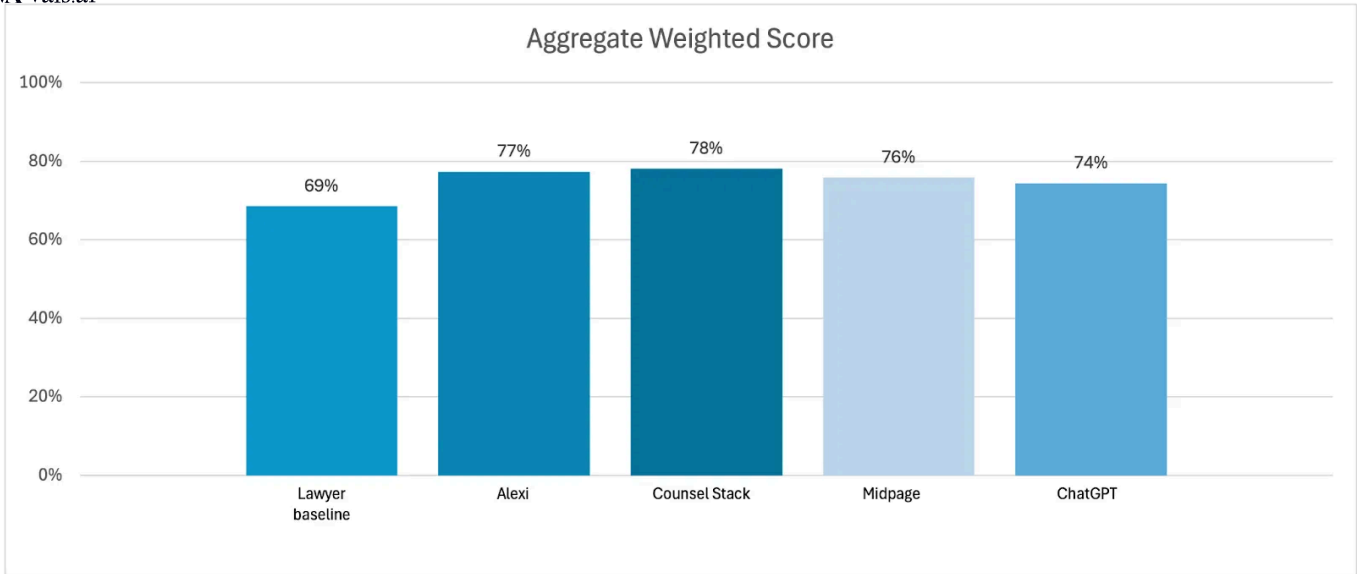
Counsel Stack received the highest score across all criteria.

Our further analysis of and conclusions based on these results are set out in the next section.

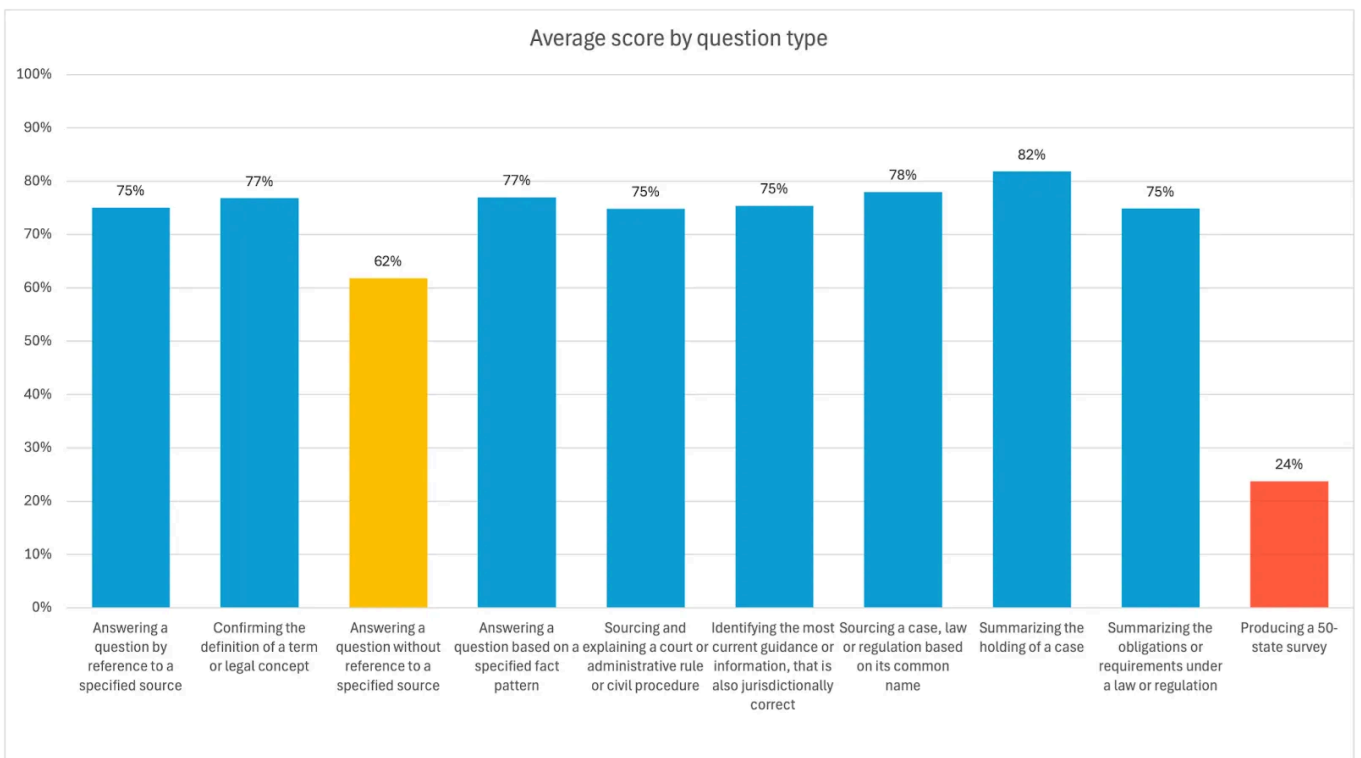
vals.ai



vals.ai



We also analyzed how the participants performed on each category of question. In general, we found that all participants performed strongly across the question types, with the exception of 50-state surveys. For the AI products, utilizing a specialist workflow rather than basic prompts should yield a much better result for this type of legal research question. There were also more mixed results where the question was vague as to, or silent on, which sources should be consulted to provide an accurate answer. This is an area where additional context is likely necessary—something that the lawyers would be able to import without requiring specific additional instruction.

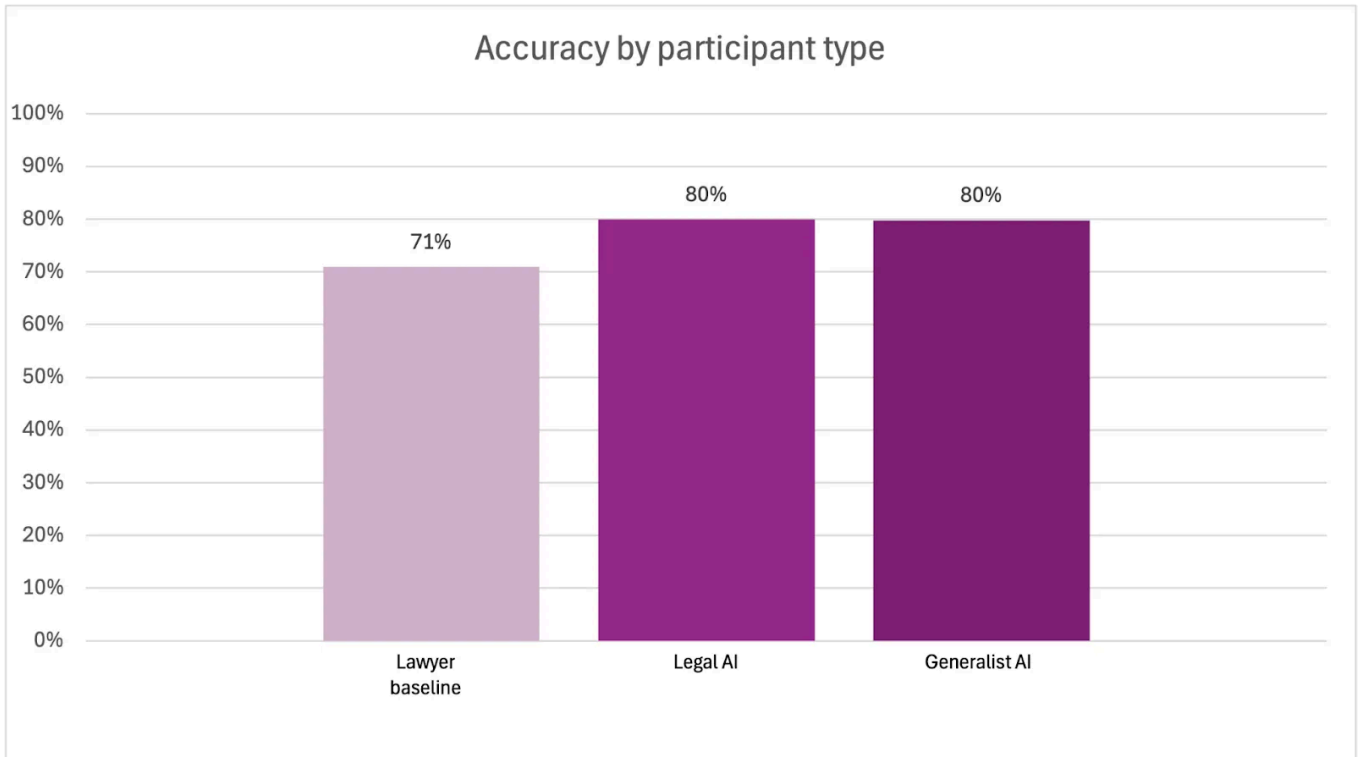


Key findings

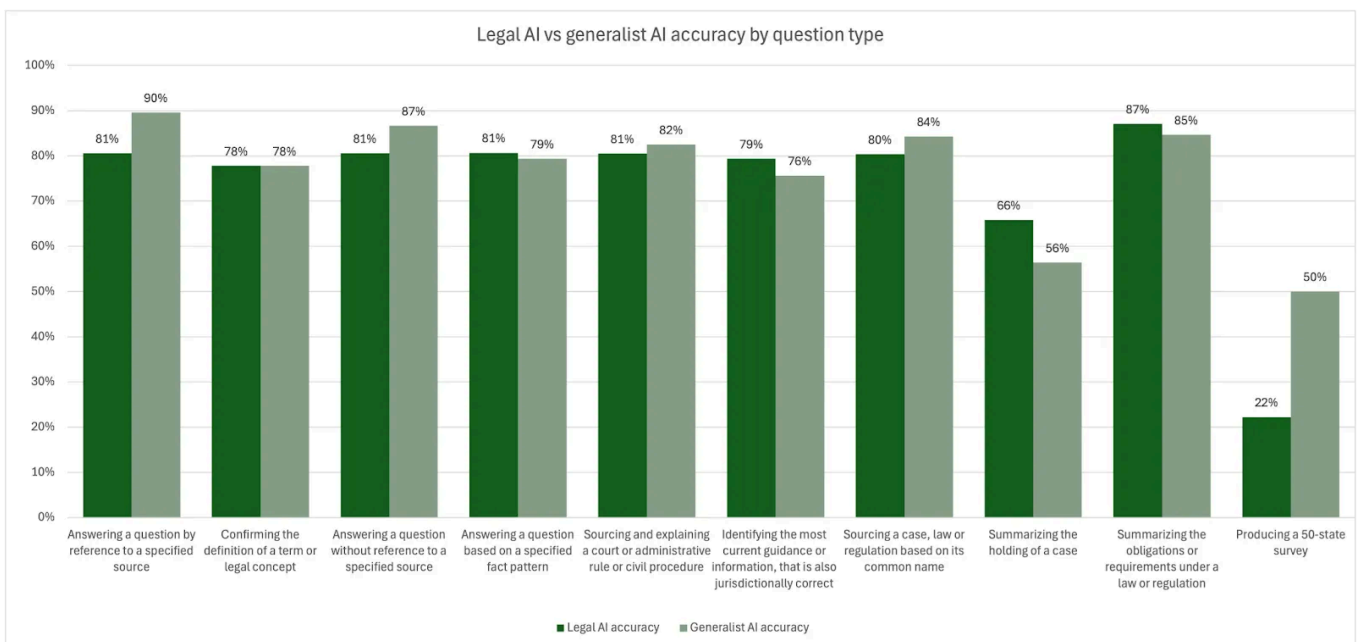
The outcomes of our study support the following key conclusions:

1. Both legal AI and generalist AI can produce highly accurate answers to legal research questions

Where the AI products were able to provide a response, overall accuracy was very strong with little differentiation between the legal AI products (78-81% accuracy, 80% averaged) and the generalist AI product (80% accuracy).



In addition, there were five question types where, on average, the generalist AI product provided a more accurate response than the legal AI products, and one question type where the accuracy was scored the same.

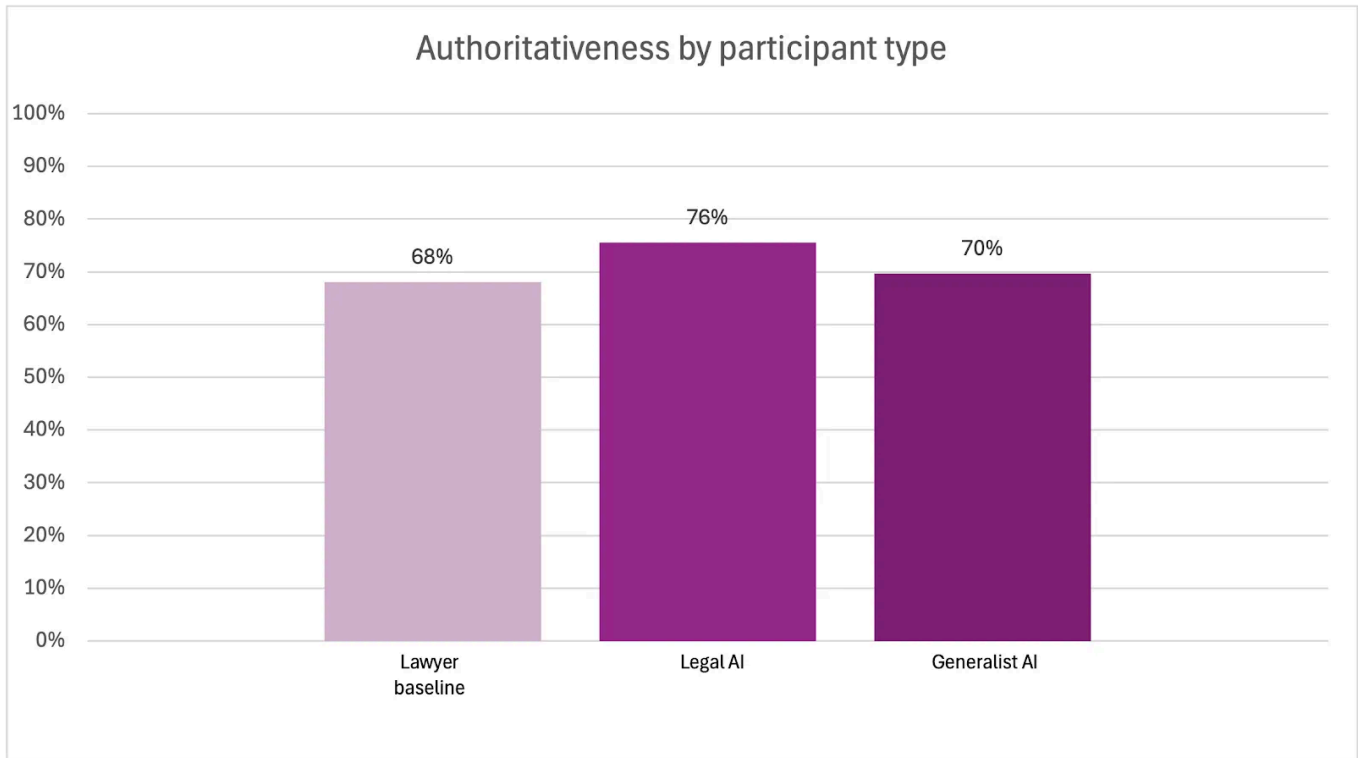


However, there were multiple instances of legal AI products being unable to produce a response due to either technical issues or deemed lack of available source data. Pure technical issues only arose with Counsel Stack (4) and Midpage (3), where no response was provided at all. In other cases, the AI products acknowledged they were unable to locate the right documents to provide a response but still provided some form of response or explanation as to why the available sources did not support their ability to provide an answer. This occurred for Alexi (2), Midpage (8) and ChatGPT (8). In some cases, the detail provided in these explanations earned the AI products partial points. This was especially true for ChatGPT which would respond that it could not find a specific judgment or law, but would still rely on its knowledge to provide an answer for the question.

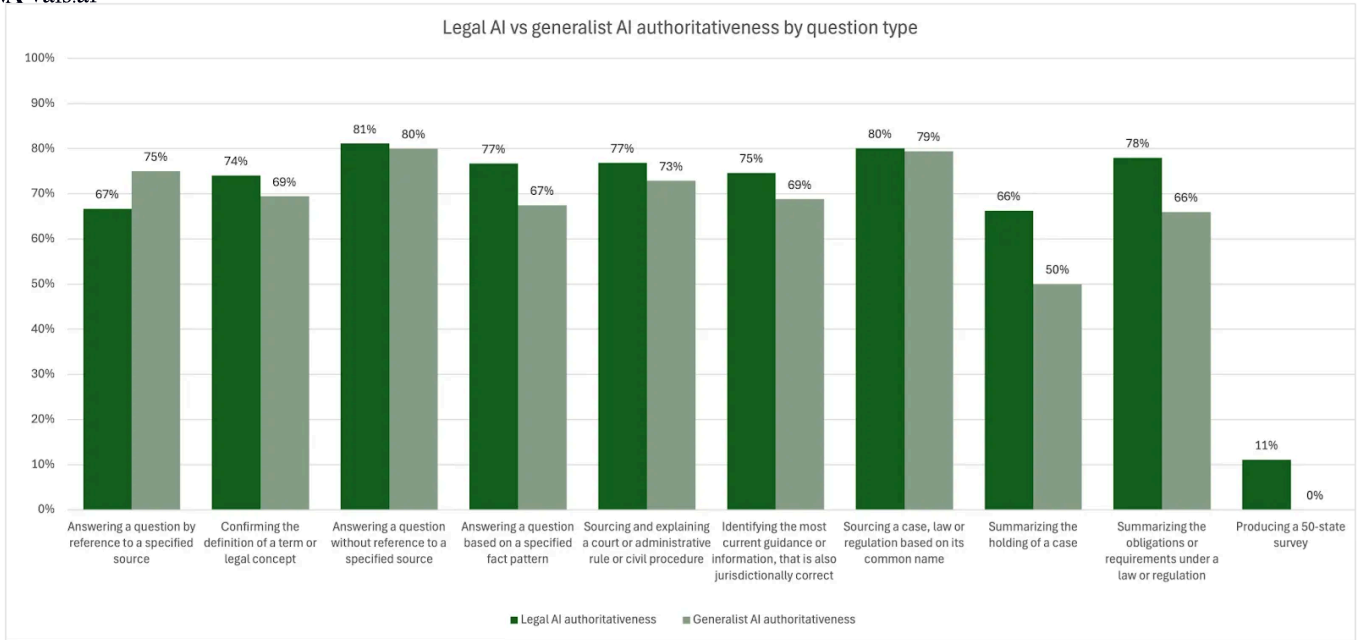
vals.ai Despite such issues, the study outcomes support a conclusion that both legal AI and generalist AI products can produce highly accurate answers to the types of legal research questions included in the study.

2. Sources and citations are the differentiators for legal AI, for now

While accuracy may not have proven to be a defining factor in the study, authoritativeness—the ability to identify and cite relevant and valid sources—provided some differentiation. The authoritativeness scores were 6 points higher on average for legal AI products versus the generalist AI product.



When looking at the authoritativeness scores across question types, there was only a single category where the generalist AI product outperformed the legal AI products. This was the category that required access to the most up-to-date guidance or information. Given ChatGPT uses web search by default, it has access to the latest information on the open web. Legal AI products often specifically restrict sourcing to their own databases to improve the reliability of their sources and citations.

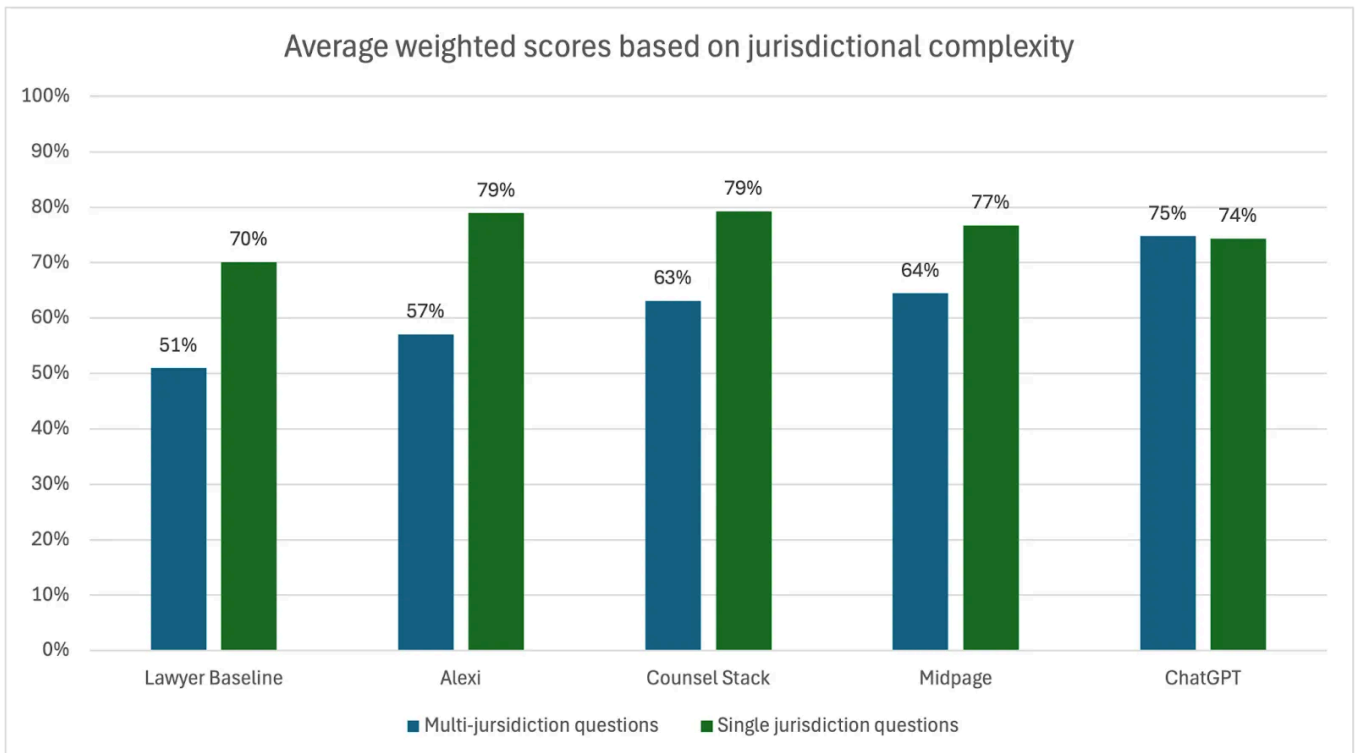


Nonetheless, the study outcomes support a common assumption that access to proprietary databases, even if composed mainly of publicly available data, does result in differentiated products.

However, this gap may close with the wide availability and adoption of Deep Research, which was announced on June 27, 2025. To our knowledge, none of the legal AI participants had implemented Deep Research in their legal research products at the time of the study response collection.

3. Most participants struggled with jurisdictional complexity

All but one participant performed significantly worse on the 14 jurisdictionally complex questions, being where the participant was asked to consider more than one jurisdiction’s laws. On average, the scores for multi-jurisdiction questions were 14 points lower than single jurisdiction questions. Only ChatGPT showed no variance based on its average scores.

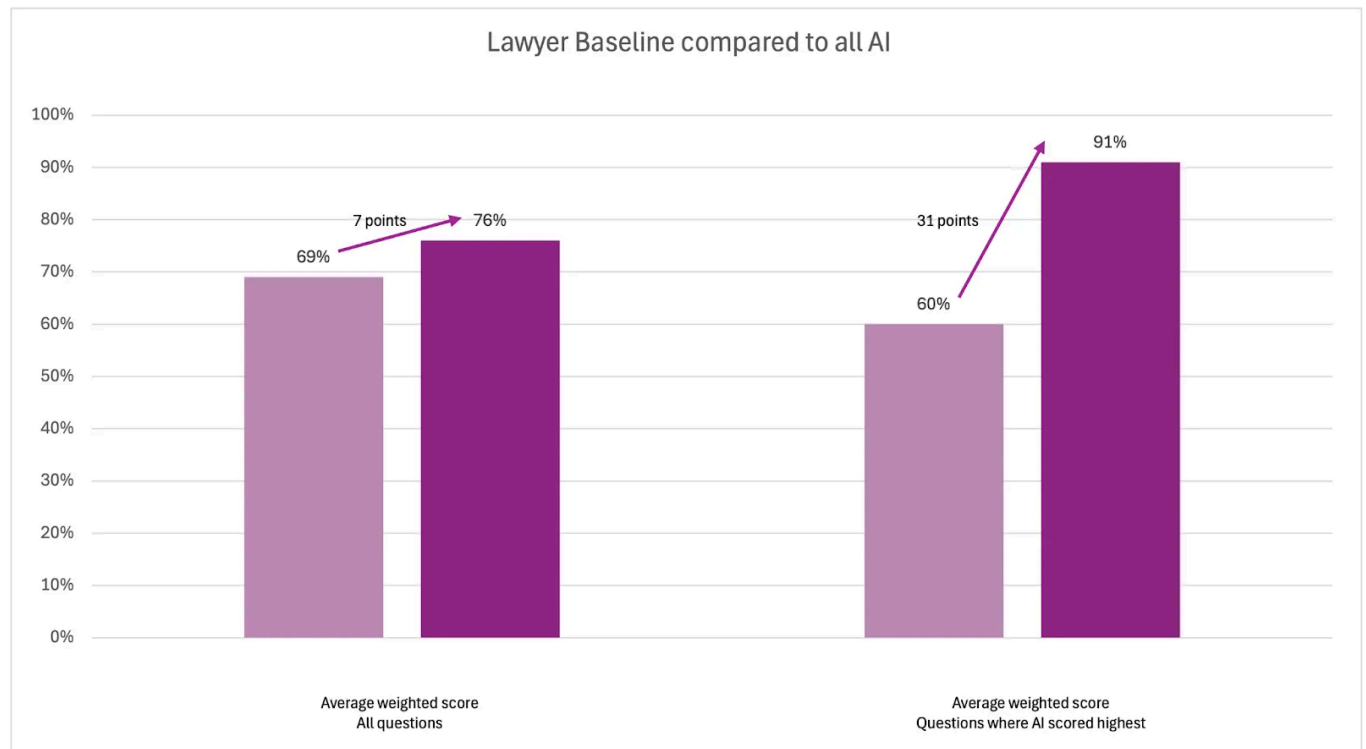


In particular, all participants struggled with the one 50-state survey question, although for different reasons. Counsel Stack failed to provide a response due to a time out error. Alexi did not attempt to source the primary law across all states and so provided only an incomplete summary from a secondary source. Both the lawyers and ChatGPT provided substantively correct and complete answers but failed to provide authoritative sources, instead primarily referencing only a handful of public sources.

For the AI products, utilizing a specialist workflow designed to search all 50 states' laws should yield a much better result for this type of legal research question.

4. Where it performs most strongly, AI can far exceed lawyer capability

The average weighted scores for the AI products exceeded the Lawyer Baseline on 150 (75%) of the study questions. Where the AI outperformed the lawyers on individual questions they did so by an average margin of 31 percentage points.

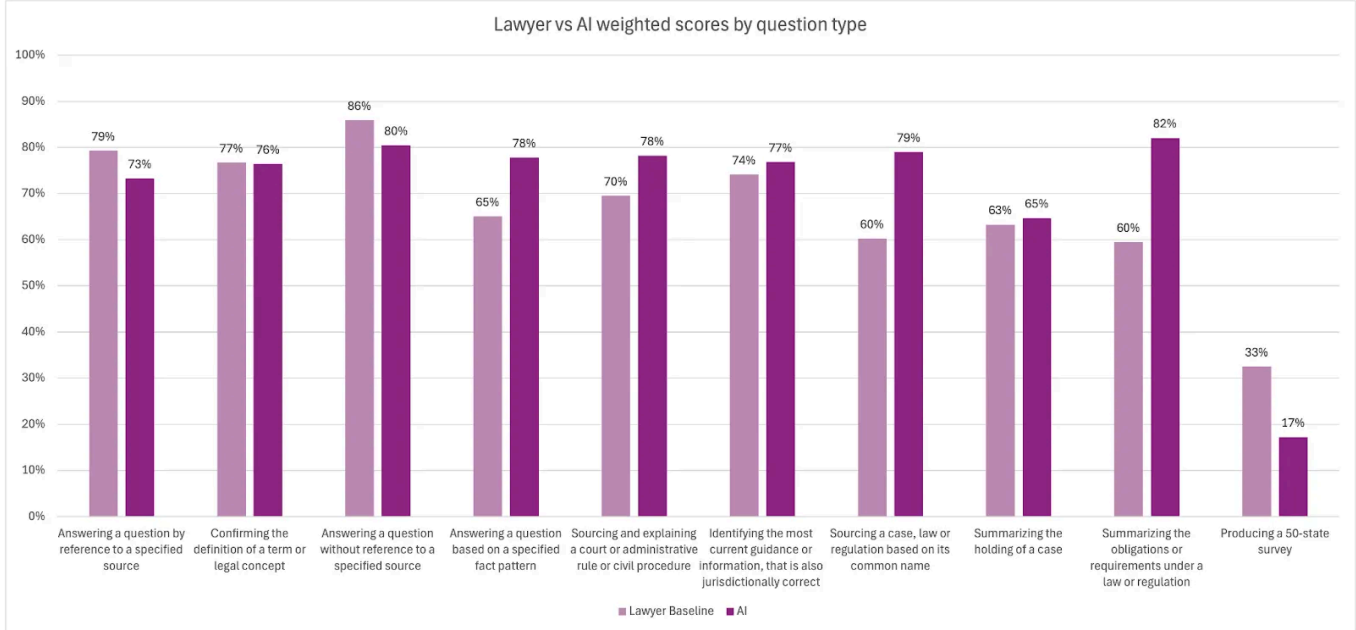


This supports a conclusion that in areas where lawyers may struggle in correctly and completely answering legal research questions, AI can provide significant support in closing the capability gap for lawyers.

5. There remain a few areas where lawyers outperform AI

For four of the ten question types, the lawyers outperformed the AI products. In addition, where the lawyers outperformed AI on individual questions they did so by an average margin of 9 percentage points. The study data suggests that the human element created an edge when a deeper understanding of context, complex multi-jurisdictional reviews and judgment-based synthesis was required.

vals.ai



It is also worth noting that the lawyers ultimately had fewer zero scores overall, meaning they often had correct responses but received only partial credit in more cases than compared to the AI products. As with the previous VLAIR study, we believe this reflects lawyers' tendency to provide more succinct, direct and (arguably) commercially-focused responses as opposed to the longer analysis that AI tends to output by default. However, this means that the lawyers' shorter responses may not have included all elements of the Consortium Firm-defined correct answers, resulting in the overall lower scores across all criteria.

Despite being numerically lower scores compared to the AI products, the Lawyer Baseline was set reasonably high (especially in the context of the first VLAIR and other benchmarking studies) and the study outcomes support a conclusion that there remain capability gaps that lawyers are best placed to fill (whether alone or, more likely, AI-assisted).

Methodology

Background

The initial VLAIR published in February 2025 intended to incorporate legal research alongside the other seven tasks covered by that study. However, in light of the increasing scrutiny of the use of AI for legal research and the expanding market of research-specific AI products, we chose to conduct this separate study to ensure a thorough and independent assessment.

Legal research remains one of the most hotly debated use cases for generative AI for a number of reasons, but especially due to the significant risk that outputs incorrectly identify or invent arguments, sources and citations.

Many in the industry continue to believe that the risk of inaccuracy and hallucination with generative AI means that legal research should only be conducted using the long-existing specialized legal research products built on top of proprietary databases of vetted sources. This view is not unfounded. There are now over 370 reported instances globally of practitioners using AI to prepare court submissions that have been found to contain hallucinated citations and/or incorrect legal analysis.

However, foundational model advancement continues at pace with capabilities like live web search and "Deep Research" now generally available to all AI product builders. While access to proprietary legal research databases remains perhaps the last remaining "moat" in the fast growing legal AI market, legal's protectionist approach seems to be giving way to collaboration. For example, many of the largest legal AI vendors have announced content partnerships in the past year.

In addition, some in the industry have openly questioned the extent to which legally-focused AI products provide sufficient value above and beyond the more accessible and cost-friendly applications built by the foundational model providers. In fact, the most repeated request we received since publishing the first VLAIR was to include generalist AI products in our future studies.

We were therefore pleased to secure OpenAI's agreement to include ChatGPT in this study. Certain of our analysis looks at the results of the three participant legal AI products against this generalist AI product in order to address the industry's calls.

Vals.ai Dataset

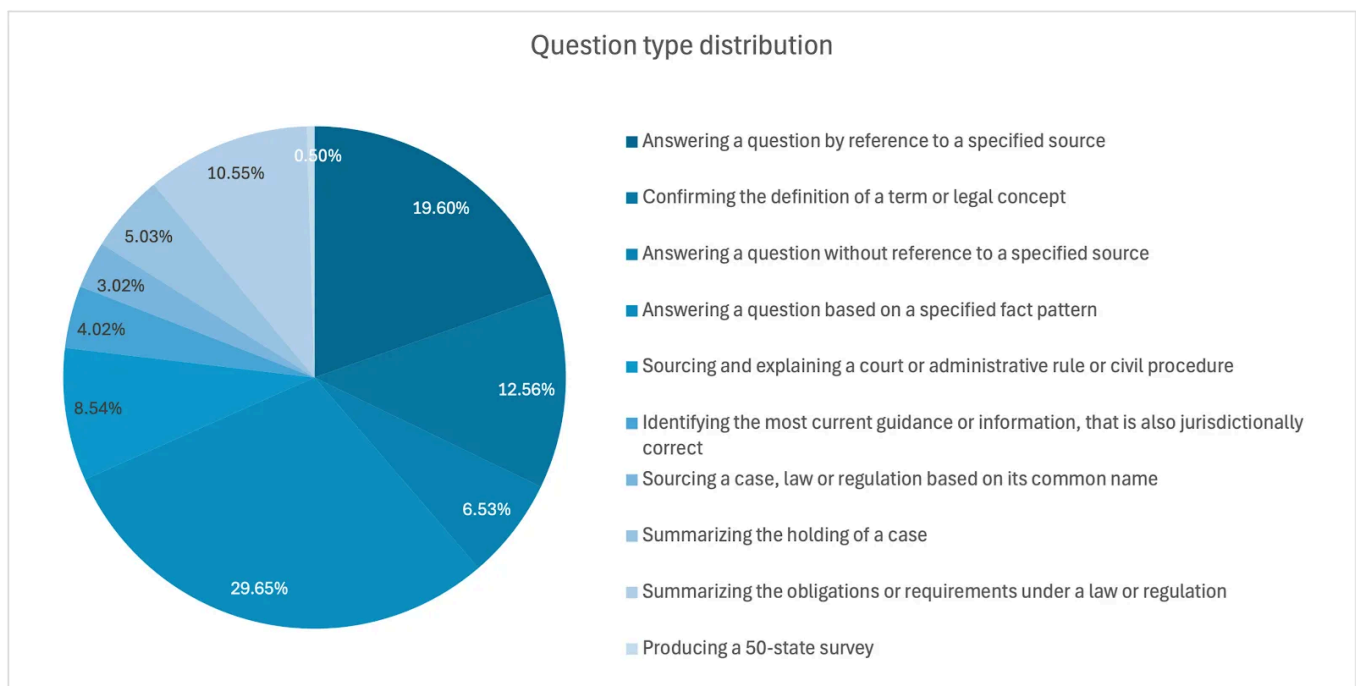
The legal research study was designed to assess the participant's ability to answer questions that require reference to U.S. federal or state laws, regulations, rules, decisions, and judgments.

In designing the legal research study, we consulted with the Consortium Firms and our practicing lawyer and law librarian network to identify the types of research questions that would be most representative of everyday private practice work.

We then asked the Consortium Firms to contribute sample questions, ideal responses including expected sources, source types (federal vs state and statute vs case law) and citations, and other indicia of correctness (objective criteria for assessing accuracy) that were distributed across the identified question types.

On completion of this process, we collected 200 question-and-answer sets (referred to as the Dataset Questions) representing one of the highest-quality legal research datasets ever assembled for studying the capabilities of generative AI tools. One question had to be disregarded when an error in the formulation was found and it was not possible to re-collect new responses from all participants.

The distribution of question types within the Dataset Questions were as follows:



A sample selection from the Dataset Questions can be accessed at the link set out in the appendix.

No vendors were supplied with a copy of the Dataset Questions.

Participating Vendors

The core tenets of the VLAIR studies are that (1) we evaluate each vendor's product with their explicit approval, (2) each product is evaluated on tasks that the product was designed to address, in the manner they were designed to be used, and (3) all participating vendors remain blind throughout the study and so cannot optimize their products for the Dataset Questions.

An open call was issued online following the publication of the first VLAIR and invitations were also issued to the most widely used legal research vendors to learn about and participate in the study. The only selection criteria a vendor was required to meet was that their product was already being used by customers for US legal research. While we had to limit the total number of participants for manageability of time and costs, this open process ensured that we had a suitably representative group of willing and active participants with whom we were able to proactively address any questions or challenges in the evaluation process.

As part of the terms of participation in the VLAIR studies, we allow each vendor to withdraw its product prior to publication. Although we had vendors opt into legal research as part of the first VLAIR, after we separated out legal research into its own study they chose not to be included.

We are fortunate to have a distribution of participants for this study comprising representatives of both specialised legal research AI providers and foundational model providers. We provide a brief introduction to the participating vendors below.

- **Alexi** is a legal AI vendor founded in 2017 that is focused on streamlining workflows for legal teams and in-house counsel. We studied their legal research product, which has been trained on case law.
- **Counsel Stack** is a legal AI vendor founded in 2023 and which offers multiple specialized LLMs for legal research, document review and client communications. We studied their legal research product.
- **Midpage** is a legal AI vendor founded in 2022 that is focused on supporting litigators with their workflows. We studied their standalone legal research product. Midpage can also be accessed via ChatGPT through a bespoke integration.
- **OpenAI** is a leading artificial intelligence research and deployment company founded in 2015. It has developed several foundational models that are used by product developers worldwide. We studied their ChatGPT Team product.

Lawyer Baseline

As with the initial VLAIR study, we chose to establish a baseline measure of the quality of work produced by the average lawyer, unaided by generative AI.

To achieve this, we partnered with a US law firm who supplied lawyers experienced in conducting legal research for client matters. The lawyers were asked to answer the Dataset Questions based on the exact same instructions and context provided to the AI products, and include any citations in-line. They were allowed to use all research resources and tools at their disposal provided they were not generative AI-based. Their responses were provided in written form within a two-week period.

The participating lawyers were also asked to report the time it took to answer each question to inform our latency comparison with the AI products.

Response collection and scoring

Response collection took place over the first three weeks of July 2025.

All questions were submitted to each product via API as a zero-shot prompt with the output returned and collected in JSON format, with one exception where no API was available. Prior to commencing the collection process, API performance was tested by the VLAIR team with each vendor to confirm it operated as expected and to ensure the output format met the study specifications. Each question was submitted at least three times to each product to reduce output variance, and questions were submitted individually in sequence and not in parallel.

The exact questions as posed in the zero-shot prompts were also provided to the lawyers who helped to create the Lawyer Baseline. No follow up questions were answered.

Due to the potential variability of correctness of any response to a legal research question, the evaluation for this study was conducted on a blind basis by a group of lawyers and law librarians and not using any automated systems.

Final collected responses were anonymized to remove details that would allow the evaluator to identify the participant, such as links to proprietary sources accessed within a participating product, before being shared with the evaluators. Links to public sources were kept within the final answer to enable checking by the evaluators.

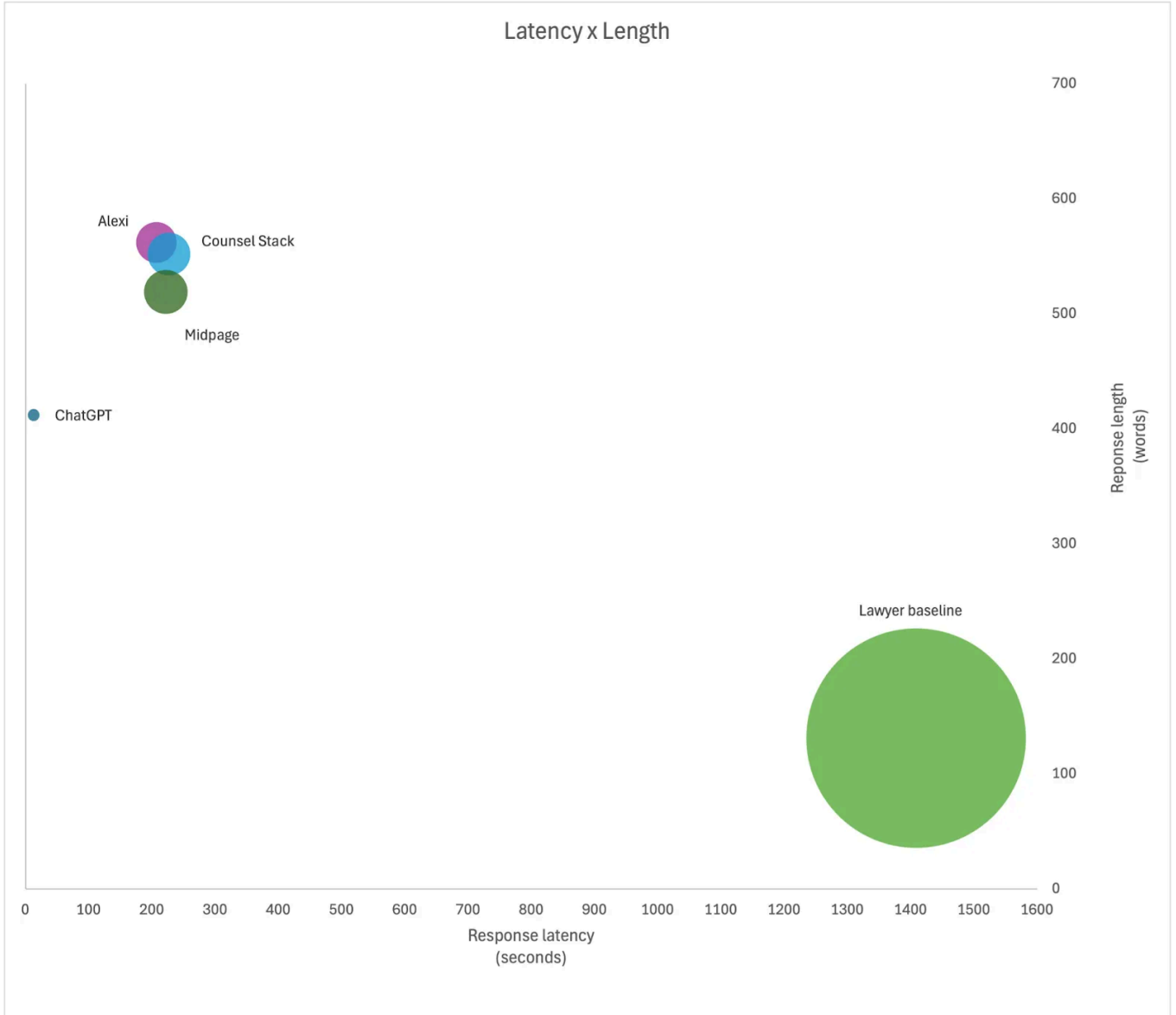
Each evaluator was assigned a specified number of questions and asked to score all of the responses to each of their assigned questions. They were asked to assess each response based on the scoring criteria set out in the appendix and considered three weighted components: accuracy (substantive correctness without incorrect elements), authoritativeness (sufficiency and strength of cited sources) and appropriateness (form, format and suitability for purpose). To support their scoring decisions, evaluators were supplied with the reference answers, sources, source types and citations, and other indicia of correctness for their assigned questions as provided by the Consortium Firms.

The responses to each question were assessed by at least two evaluators. Once the scores of each evaluator were submitted to the Vals study team, they were averaged to produce a final score for each criteria, with zero scores escalated to a third evaluator for verification. Once the zero score verification process was completed, weighting was applied to each criteria to produce the overall average weighted score for each response.

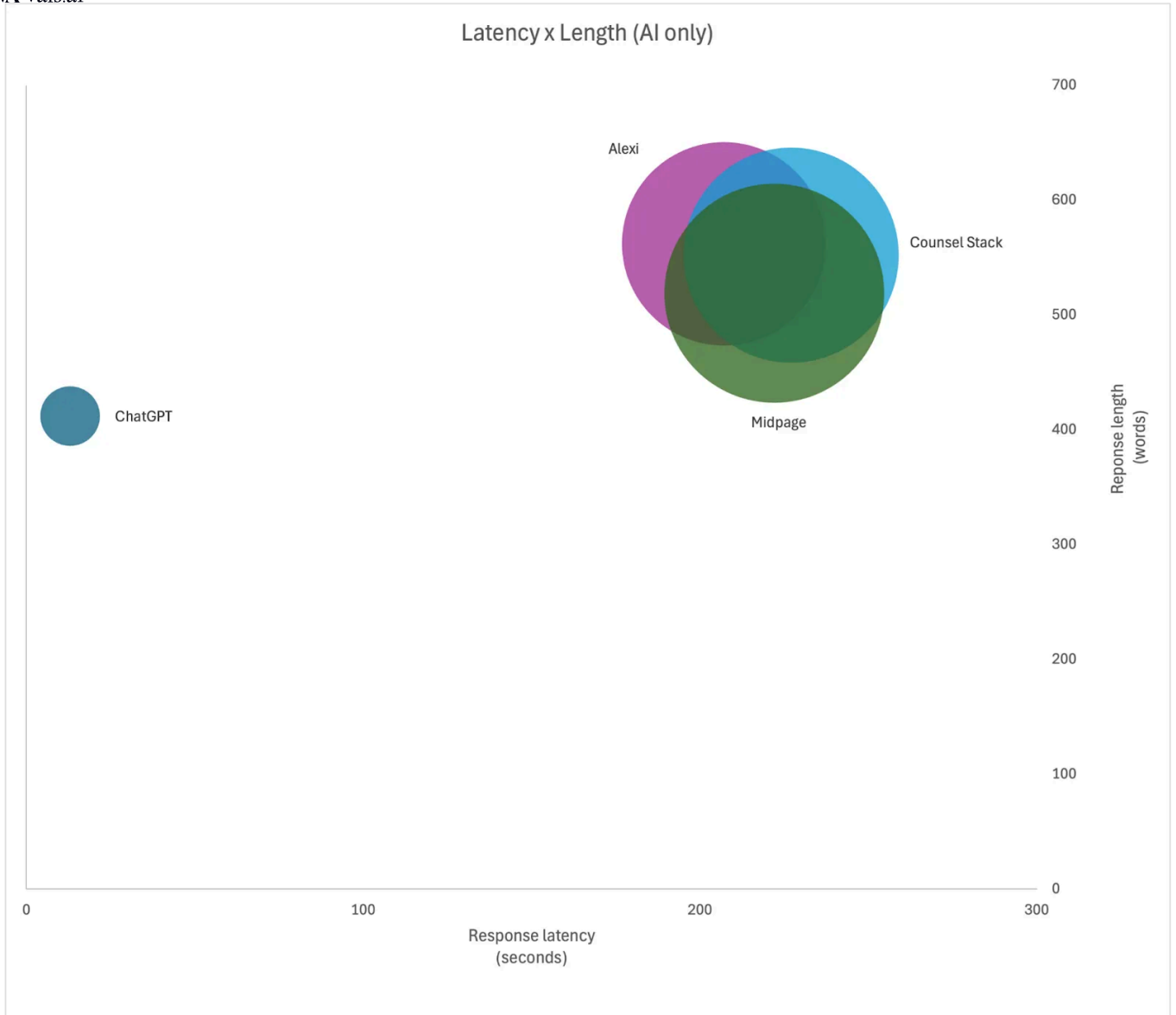
Additional findings

vals.ai Latency x Length

We measured the speed with which the participants could return a response (latency) and contrasted this with the length of responses. The total time to return a response was calculated in seconds, starting as soon as the question was submitted and completing when every token was returned. The length of response was calculated as the total number of words returned by each participant. When both of these metrics were averaged for each participant we were able to calculate the amount of thinking and generation time per word, which is represented by the bubble size in the below chart.



vals.ai



Limitations

Legal research encompasses a wide range of activities calling upon various capabilities to understand context, identify the most appropriate jurisdiction and source type, locate the best supporting sources and provide precise citations to such sources so others can find and confirm the response is correct and sufficiently supported.

As a result, there is not always a 100% correct answer that can be provided in advance. This is especially the case in the context of providing the strongest and most relevant sources. For some questions, there could be an almost limitless number of potential sources, whereas for others there may be a single correct source. Correct answers may also change over time as laws, regulations or precedents also change.

All of this is what makes legal research a challenging area for lawyers and AI alike, and makes benchmarking AI products on legal research questions an equally challenging endeavor.

We acknowledge certain limitations that may make the results of this study more or less useful for some when conducting their own evaluations of AI products for legal research, including the following.

- Under the methodology for this study, we asked the Consortium Firms to provide a minimum correct reference answer while also giving them the opportunity to provide other indicia of correctness for evaluators to use when scoring. This was so that the evaluators did not themselves need to research the correct answers, which would have significantly elongated the evaluation process. Our evaluators were allowed to exercise discretion where they felt the reference answer had minor errors or was less complete than ideal, so that correct and more complete responses would not be penalized.

- Because some of the legal AI products have filtering options, we included in the question context an indicator of which jurisdiction(s) should be searched. For example, a question asking to summarize the holding of a named U.S. Supreme Court decision would include an instruction to “Search federal case law to answer this question.” As a result, the true sourcing capability of the AI products may not have been tested in full. In addition, after the response collection process was completed and the evaluation process commenced, we found two instances of a question including an incorrect jurisdictional reference. We were able to re-run one of these corrected questions, with new outputs collected from all participants. The second instance was only flagged at the end of the evaluation process after which API access had been revoked by the vendors. We therefore decided to disregard this question for the study, although we note that 3 of the 4 AI products, as well as the lawyers, correctly identified the error in the question and provided an answer based on the correct jurisdictional laws. While this is an interesting result, because the misdirection was not intended we felt it was most fair to disregard the question all together.
- Posing legal research questions as a simple zero-shot prompt is not necessarily true-to-life in terms of how a lawyer or research professional seeks to find an answer to a legal research question. While we sought to ensure that the study reflected the diversity of types of questions that are most asked, we did not cater for follow-up prompting or addition of context, nor did we use any workflow-based features of any AI product. Moreover, there were certain research-related tasks, such as specifically sourcing materials to support particular arguments or to generate specifically formatted citations, that the study did not cover.

Acknowledgments

This report would not have been possible without the support and contributions of:

- The contributing law firms (referred to as the Consortium Firms), including Reed Smith, Fisher Philips, McDermott Will & Emery, Ogletree Deakins, Paul Hastings and Paul Weiss. The Consortium Firms helped to create the dataset used for the study and provided additional guidance and input to ensure the study best reflected real-world legal research questions and answers.
 - The participant vendors, Alexi, Counsel Stack, Midpage and OpenAI, who agreed to be evaluated and provided access to and assistance with their AI tools. Note that Vals AI has a customer relationship with one or more of the participants. We respect their effort to build an unbiased study of legal research products.
 - The US lawyers who participated in the study to create the Lawyer Baseline.
 - The lawyer and law librarian evaluators, including: Anna Guo and Mohamed Al Mamari both of LegalBenchmarks.ai, T. Kyle Turner and Emily Pavuluri both of the Vanderbilt AI Law Lab and Nick Hafen of BYU Law. Additional evaluators have elected to remain anonymous.
-

Appendix

Sample Dataset

A set of example questions have been provided below. More samples can be accessed by getting in touch with contact@vals.ai.

Example Q&A Scoring

To help illustrate and contextualize the study results, particularly as to the performance by the different types of participants, we include below some example scoring from questions in the Sample Dataset.

Example 1 - Where lawyers (100%) significantly outperformed all AI (3% to 80%)

Contextual understanding of legal market terms

Dataset Q&A	Commentary
<p>Q: Is there a California provision that allows a right of rescission or treble damages if a sponsor sells through an unlicensed broker?</p> <p>Search California primary law to answer this question.</p> <p>A: California Corporations Code section 25501.5 allows a person who purchases a security from or sells a security to an unlicensed broker-dealer to bring an action for rescission of the sale or purchase, or for damages if the security is no longer</p>	<p>A correct response needed to understand that the references to a “sponsor” and “broker” were to a financial sponsor and broker-dealer—making the subject matter of the question the sale of securities—and reference the civil liability provisions of the California</p>

vals.ai Dataset Q&A

Commentary

owned. Rescission action against unlicensed broker–dealer. This statute provides a clear remedy for transactions involving unlicensed broker-dealers, ensuring that the injured party can recover their losses.

Corporations Code which covers both the right of rescission and treble damages.

California Corporations Code - Title 4 Securities - Division 1 Corporate Securities Law of 1968 - Part 6 Enforcement - Chapter 1 Civil Liability - Section 25501.5
<https://codes.findlaw.com/ca/corporations-code/corp-sect-25501-5.html>

The lawyers understood the context of the question, provided a correct answer and cited the right provisions of California law.

The AI products either misinterpreted the question as relating to real estate and/or failed to identify the correct provisions of California law. Scores as between legal AI and generalist AI were mixed.

Example 2 - Where all AI (80% to 100%) significantly outperformed the lawyers (18%)

Ensuring a sufficiently detailed response

Dataset Q&A

Commentary

Q: Does the Federal Trade Commission regulate or have jurisdiction over nonprofit organizations or trade associations?

A correct response needed to identify that the general position is that the FTC does not exercise jurisdiction over non-profits and trade associations, and then explain the limited exceptions where there is certain for-profit activity and that such activity may change the non-profits' definition under the FTC Act. The best supporting authorities were the FTC Act itself and supporting case law.

Search federal primary law and case law to answer this question.

A: The Federal Trade Commission does not generally exercise jurisdiction over nonprofit corporations or trade associations that are organized and engaged in business for purely charitable purposes, as specified in Section 4 of the FTC ACT, 15 USC Sections 41-45. The FTC may have jurisdiction over trade associations by virtue of a trade association's status as nonprofit if the association conducts business for the sake of their members, or when an association engages in lobbying and litigation activity with respect to matters affecting members' pecuniary interests. Nonprofits can be considered a "person" or "partnership" under the FTC Act if it engages in non-profit oriented activities or provides significant economic benefits to its members.

Federal Trade Commission Act, 15 USC 41-45
<https://www.law.cornell.edu/uscode/text/15/45>
 Cal. Dental Ass'n. v. FTC, 526 U.S. 756 (1999)
<https://supreme.justia.com/cases/federal/us/526/756/>

FTC v. National Com. On Egg Nutrition, 517 F2d 485 (7th Cir. 1975)
<https://law.justia.com/cases/federal/appellate-courts/F2/517/485/155891/>

The AI products all provided detailed explanations of the types of entities within the jurisdiction of the FTC, how certain activities affect the status of non-profits and trade associations and the decisions from relevant case law that helped to establish those exceptions.

The lawyers focused only on the jurisdictional definition within the FTC Act and a single court ruling without significant elaboration on how the case was relevant. The response did not provide particular analysis explaining the exceptions and the cited case was not authoritative.

Example 3 - Where legal AI (91% to 100%) significantly outperformed generalist AI (8%)

Identifying and accessing particular legal sources

vals.ai

Dataset Q&A

Commentary

Q: Where can I find the 2025-2026 campaign contribution limits and what are the limits for donating to political action committees?

Search federal primary law to answer this question.

A: The 2025-2026 contribution limits were published in the Federal Register on January 30, 2025 at 90 Fed. Reg. 8526.

Any donor type can contribute up to \$5,000 per year to a political action committee.

90 Fed. Reg. 8526 (January 30, 2025)

<https://www.fec.gov/resources/cms-content/documents/fedreg-notice-2025-01.pdf>

A correct response needed knowledge that the contribution limits are published annually in the Federal Register and then precisely source the information for 2025-2026.

The legal AI products all answered correctly although a few failed to cite the Federal Register, instead pointing to the CFR and/or other public sources.

The generalist AI product failed to attempt to respond, instead suggesting the user check the Federal Election Commission website or sourcing other official legal resources. Still, an average half point was awarded by the evaluators for unknown reasons.

It is worth noting that the lawyers received a 100% on this question.

Example 4 - Where generalist AI (82%) significantly outperformed legal AI (0% to 30%)

Sourcing up-to-the minute information

Dataset Q&A

Commentary

Q: How many U.S. states have laws or regulations that require the use of uniform applications for hospital-based financial assistance policies (FAPs)?

Search the primary law of all 50 states to answer this question.

A: As of March 2025, three states (Colorado, Maryland, and New York) have laws or regulations that require the use of uniform applications for hospital financial assistance policies. Colorado passed legislation in 2024 to repeal the Colorado Indigent Care Program (CICP), which was Colorado's uniform application for hospital financial assistance policy. CICP remains effective until July 1, 2025.

2024 Colorado Revised Statutes. Title 25.5-Health Care Policy and Financing, Article 3-Indigent Care, Part 5-Health Care Billing for Indigent Patients Receiving Services Not Reimbursed Through the Colorado Indigent Care Program, Colorado Revised Statutes § 25.5-3-502

<https://law.justia.com/codes/colorado/title-25-5/indigent-care/article-3/part-5/section-25-5-3-502/>

Annotated Code of Maryland, Title 19. Health Care Facilities, Subtitle 2. Health Services Cost Review Commission, Part II. Health Care Facility Rate Setting, Section 19-214.1, Financial Assistance Policies, Md. Code, Health-General, §19-214.1

<https://law.justia.com/codes/maryland/2005/ghg/19-214.1.html>

Consolidated Laws of New York, Chapter 45, Public Health, Article 28, Hospitals, Section 2807-K, General Hospital Indigent Care Pool, N.Y. Public Health Laws, § 2807-K

<https://www.nysenate.gov/legislation/Laws/PBH/2807-K>

A correct response should have identified the three states that use uniform applications for FAPs, noting that Colorado repealed their law effective July 1, 2025. Significant additional detail was not requested.

The generalist AI product correctly identified that two of the three states had effective laws concerning FAP (with the data collection having taken place post July 1st), but did not cite to any primary law courses, only an article covering the topic.

The legal AI products either failed to identify any laws or failed to identify all three of the states that had laws. None identified Colorado and, therefore, recognition that its law had recently been repealed was not included in their responses.

It is worth noting that the lawyers received a 55% on this question, evidencing the challenge of searching across all 50 states' laws.

Example 5 - Where all participants responded correctly (98% to 100%)

Pinpointing data by reference to a specified source

Dataset Q&A

Commentary

Q: What are the new threshold amounts for the securities portfolio test and the initial margin and premium test under the updated Portfolio Requirement in CFTC Rule 4.7(a)(1)(v)?

A correct response needed to discern that the cited CFTC rule has been recently updated and identify the rule change as

Search federal primary law to answer this question.

would be published in the Federal Register

Join our mailing list to receive benchmark updates

Model benchmarks are seriously lacking. With Vals AI, we report how language models perform on the industry-specific tasks where they will be used.

Subscribe

By subscribing, I agree to Vals' Privacy Policy.

Copyright © 2025 Vals AI, Inc. All rights reserved.

[X \(Twitter\)](#) [LinkedIn](#)

- [Benchmarks](#)
- [Methodology](#)
- [Models](#)
- [Platform](#)
- [About us](#)
- [Updates](#)
- [News](#)
- [Careers](#)
- [Privacy Policy](#)

1 = The response is substantively correct, but contains misinterpretations, factual errors and/or omissions that are significant enough to call into question its reliability as a whole.

0 = No response was given or the response is substantively incorrect.

<p>Authoritativeness</p> <p>Whether the response cites relevant primary sources (i.e. from primary law or an appropriate court that supports the statements) that are valid (i.e. exists and remains good law) and, at a minimum, includes those citations referenced in the reference answer.</p>	<p>Up to 3 points can be awarded:</p> <p>3 = The citations include those set out in the reference answer with only minor errors in form (name, reporter, court, year), and may include additional relevant and valid sources.</p> <p>2 = The citations provided are relevant and valid, but are not as strong as those included in the reference answer, for example because they refer to a lower court when the matter was also decided at a higher court or come from a non-binding jurisdiction, or a draft bill or non-enacted legislation rather than the adopted primary law or include law that has been amended or superseded.</p> <p>1 = The citations provided include some of those in the reference answer but also (a) erroneous citations have been included that do not support the answer, or (b) statement(s) lack a citation where clearly one should be provided.</p> <p>0 = Either: (a) no citations are provided at all, or (b) citations are provided but they neither include those in the reference answer nor otherwise support the answer.</p>	<p>40%</p>
<p>Appropriateness</p> <p>Whether the response is easy to follow and understand, and does not contain grammatical or spelling errors. It could be immediately shareable with colleagues or clients.</p>	<p>Up to 2 points can be awarded:</p> <p>2 = The response is clear and easy to understand with no errors, and could be immediately shared.</p> <p>1 = The response is (a) either overly terse or overly verbose (having reference to what the question is asking), and/or (b) the response contains errors or deficiencies in its presentation, clarity or form, in each case that would need at least 15 minutes of correction</p>	<p>10%</p>

Criteria	Description	Scoring and guidance	Weight
vals.ai		before sharing. 0 = No response was given or the response is unusable as it would need a significant rewrite before sharing.	