Who Checks the Citations? Benchmarking Legal Hallucination Detection

Patty Liu,[†] Dominik Stammbach, Peter Henderson[†]
Princeton University

Dataset | Code | Website

Attorneys, judges, and pro se filers increasingly use AI to draft legal documents, yet these tools frequently fabricate citations. Despite predictions that newer models would hallucinate less or that court sanctions would deter negligent filers, we found over 1,000 filings containing fabricated citations—with this number growing year-over-year. This study evaluates whether AI-based systems can mitigate these errors by automatically detecting hallucinations. We propose a taxonomy of legal citation hallucinations grounded in actual court filings and introduce a dataset of 1,300 brief excerpts containing injected errors. Benchmarking five models in agentic and non-agentic settings reveals that while the latest iterations perform better—GPT-5 achieves 82.8% recall and a 60.5% F1 score in an agentic framework—all models struggle with subtle error categories. Agentic verification remains resource-intensive, with GPT-5 averaging 16.9 steps per excerpt. Furthermore, restricted information access limits the efficacy of even the best agents. This gap creates policy concerns, as it disadvantages both AI systems and litigants who lack subscriptions to commercial legal databases. Together, our dataset, tools, and policy recommendations provide a foundation for building and auditing reliable legal citation checking tools.

1. Introduction

Legal citations play a prominent role in U.S. legal practice. Attorneys must point to past judicial decisions and laws to make their case (Duxbury, 2008; Lamond, 2016). Fabricating citations, or misrepresenting the content of those citations, is the same as pointing to made-up law to win the case—courts have called it "an abuse of the adversary system" that risks the integrity of the judicial process. *Mata v. Avianca, Inc.*, 678 F.Supp.3d 443 (S.D.N.Y. 2023); *see also Park v. Kim*, 91 F.4th 610 (2d Cir. 2024) (quoting *Mata*); *Noland v. Land of the Free*, 114 Cal.App.5th 426, 445 (2025) (quoting *Mata*).

In the past, such fabrications existed but were rare.[1] The rapid adoption of large language models (LLMs) in the legal system, though, has turned such rare individual instances into a systemic problem. Pro se litigants,[2] trained attorneys, and even judges are using LLMs to generate briefs, motions, and other court filings. But even the best LLMs still fabricate or misrepresent legal precedent at a non-negligible rate (See Figure 1 and § 2). Combining our own research with other public datasets, we identify over 1,000 court filings containing hallucinated citations.[3] Judges repeatedly describe the resulting burden as an "enormous waste of judicial resources," and increasingly impose sanctions because "lesser sanctions have been insufficient to deter the conduct." *Mid Cent. Operating Eng'rs*

[†]Correspondence to: patty.liu@princeton.edu; peter.henderson@princeton.edu.

[1] *See, e.g., Gonzalez-Ayala v. United States*, No. 3:05-cv-01291 (D.P.R. Dec. 2, 2009), ECF No. 8 (discussing such a made-up citation).

[2] Litigants who represent themselves without an attorney.

[3] See project website for our compilation of tracked cases, which is merged with additional data from Charlotin (2026).

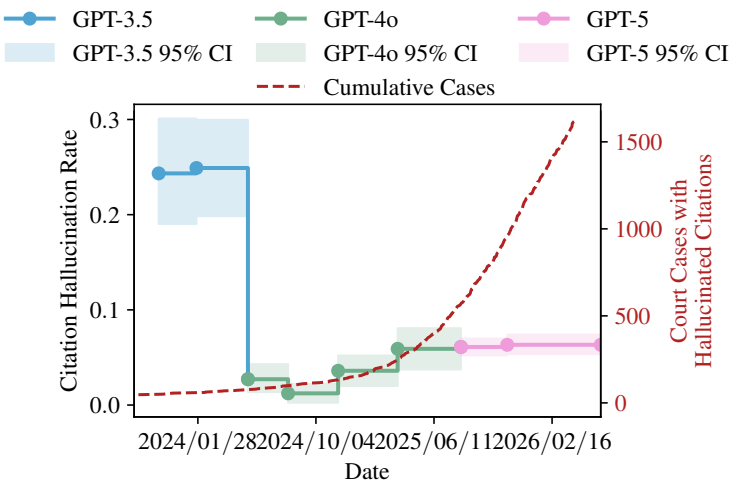


Fig. 1: **Hallucination trend in legal filings and model hallucination rates over time.** Legal hallucination rates are not consistently decreasing across GPT models, while hallucinated citations in court filings are steadily growing. Hallucination rate is the percentage of hallucinated citations in all model generated citations (See § 2 for more details).

Health & Welfare Fund v. HoosierVac LLC, 2025 WL 574234, at *3 (S.D.Ind. Feb. 21, 2025); *Powhatan County Sch. Bd. v. Skinger*, 2025 WL 1559593, at *10 (E.D.Va. June 2, 2025).

A prevailing argument has been that this problem is temporary, that hallucination rates would diminish as models improve, and that high-profile sanctions would induce greater caution. Our findings and recent events challenge these assumptions. Real-world court filings containing hallucinated citations are growing steadily, and the problem is not yet self-correcting (see Figure 1). Through a controlled experiment querying eight generations of ChatGPT models on 92 legal drafting prompts, we find that hallucination rates are no longer consistently decreasing across model generations, with GPT-5.1 producing hallucinated citations at a significantly higher rate than the best mid-2024 GPT-4o release ($p = 0.001$). Even if hallucination rates improved, the verification burden would continue to grow. Newer models generate more citations per document than their predecessors, drawn from a broader and less canonical set of cases that are individually harder to verify (Figure 2). The result mirrors the dynamic that Chang (2026) calls legal practice's own Jevons paradox: as AI reduces the per-unit cost of legal drafting, total filing volume is likely to grow, increasing verification burdens even if individual hallucination rates improve.[4]

We argue that automated verification tools, improving open, public access to legal data, and targeted AI literacy guidance for unrepresented filers offer more structural paths forward (§ 5). This work focuses on the first: currently, verifying case citations is largely a manual process performed by attorneys, law clerks, or judges, and especially when a cited case does not exist, this can take significant time and resources. Yet no prior benchmark evaluates how well AI systems can verify legal citations—instead, most research focuses on whether LLMs produce factually correct legal explanations, case summaries, or answers to legal questions (Dahl et al., 2024; Deroy et al., 2023; Fan et al., 2026; Feijo and Moreira, 2021; Hu et al., 2025; Savelka et al., 2023) (§ 6).

To assess the promise of AI in automatically checking legal filings, we introduce LEPHANTOM-

[4]See ABA Law Practice Magazine (2025); Artificial Lawyer (2024); Jevons (1865) for more information on Jevons paradox and its application to legal AI.

CITE,[5] a new benchmarking dataset of legal brief excerpts augmented with injected hallucinations. The dataset enables us to benchmark citation verification systems and is accompanied by a taxonomy of legal citation hallucinations derived from failure modes observed in real court filings. Using this dataset, we evaluate several LLMs and an agentic system based on Zheng et al. (2026) that integrates search tools and structured reasoning. Our best agent reliably detects non-existent cases and case name mismatches, but struggles with verifying pincites,[6] misquotes, and content misrepresentations. The agent also surfaced several citation errors in pre-LLM briefs submitted to state supreme courts, suggesting practical value beyond the benchmark.

Taken together, we make the following contributions:

- We provide longitudinal empirical evidence, across eight ChatGPT models from late 2023 through late 2025, that legal citation hallucination rates do not consistently decline, while the number of hallucinated court filings grows steadily.
- We introduce a dataset of 1,300 legal brief excerpts with injected hallucinations to support evaluation and auditing of citation verification systems, grounded in a taxonomy of legal citation hallucinations derived from real court filings.
- We evaluate several different models using a custom harness giving access to legal database searches and information-directed exploration. Agentic retrieval improves recall by 19.7% over non-agentic baselines on GPT-5, but all models struggle most with incorrect pincites, verbatim misquotes, and content misrepresentation (18.2%, 82.6%, and 84.0% recall for GPT-5 respectively).
- We identify potential policy options for improving the state of automated agentic citation checking. Some failures were due to lack of easy access to publicly available data on legal citations or page number information, making it especially challenging to verify this information. Improving data accessibility for the general public would, in turn, reduce burdens on courts.

Overall, we hope to incentivize improvements in automated verification to ease increasing burdens on the legal system. Our benchmarking effort presents a first step toward that goal.

2. Hallucination Rate Trends and the Growing Verification Burden

To demonstrate the pervasive challenge of hallucinations in legal content, we conduct a controlled experiment to assess whether newer and better models have reduced citation hallucination rates. We query eight generations of ChatGPT models,[7] spanning from the first GPT-3.5 model through GPT-5.1, on the same 92 prompts designed to generate legal documents. The prompts ask the LLM to generate documents similar to the ones flagged in real court filings because of AI-generated hallucinations (examples in Appendix Table A3), producing over 8,000 case citations in total. We verify all citations using CourtListener and Westlaw.[8] We classify a citation as hallucinated if it does not correspond to a real case or if the generated case name does not match the name associated with the reporter reference in Westlaw. For example, the citation *995 F.2d 348* in *"In re Grand Jury Subpoenas, 995 F.2d 348, 352–53 (2d Cir. 1993)"* exists, but refers to the case *Malcolm v. National Gypsum Co.* We classify it as hallucinated due to the inconsistency. This definition captures only the most obvious

[5] Short for Legal Phantom Citation.

[6] A pincite specifies the exact page or paragraph within a source where the cited material appears.

[7] All eight models were the default model for free-tier ChatGPT access in the browser interface at the respective times, likely to be used by pro se litigants and even many attorneys.

[8] CourtListener is a free, open-access repository of U.S. court opinions maintained by the nonprofit Free Law Project and Westlaw is a subscription-based legal research platform by Thomson Reuters.

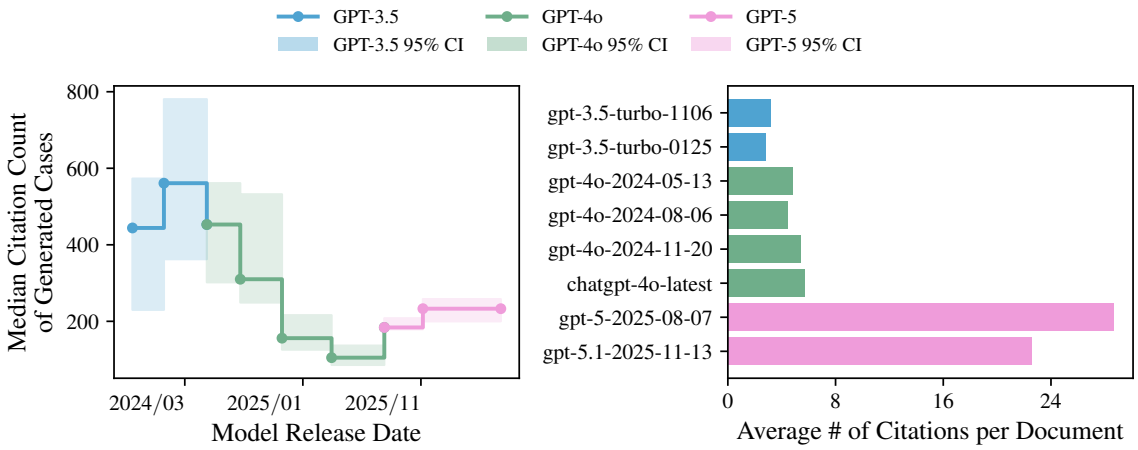


Fig. 2: **Citation counts of the generated cases (left) and average number of citations per document (right).** GPT-4o models generate citations to the most diverse set of cases, with a median of 105 citations per generated case (left), whereas ChatGPT3.5 generated case citations to mostly widely known landmark cases with a median citation number of over 500. Newer models produce more legal citations per document (right).

hallucinations and provides a conservative lower bound on overall prevalence of such hallucinations. Subtler misrepresentations, such as citing a real case for a non-supported proposition, are not captured by this criterion.

Figure 1 shows the cumulative number of identified court filings containing hallucinated citations and hallucination rates observed across models in our experiment. Early GPT-4o models released in mid-2024 exhibit the lowest hallucination rates at 1.23%, substantially improving over citation rates of around 25% in earlier GPT-3.5 models. However, this trend does not persist and more recent models hallucinate legal citations more often again. The GPT-5.1 model generates hallucinated citations at 6.57%, a higher rate than the August 2024 GPT-4o release (using bootstrapping, p$p = 0.001$).

The hallucination rate alone does not fully capture the growing verification burden. Newer models also generate substantially more citations per document than their predecessors, and draw these citations from a broader set of cases: GPT-3.5 models mostly cite well-known landmark cases with a median citation count of over 500, compared to GPT-4o citing less canonical cases with a median citation count of 105 at the lowest. While greater citation diversity may produce better and more tailored legal arguments, lesser-known cases are harder to verify (See Figure 2).

The citation verification burden affects not just filers. When filers do not verify citations, the task shifts to opposing counsel, law clerks, and judges, whom courts agree are ill-positioned to handle it. See, e.g., *Nelson v. Navient Solutions, LLC*, 2025 WL 2633962, at *3 (S.D. Iowa Sept. 4, 2025); *Fang v. Hechalou US LLC*, 2025 WL 3049873, at *3 (C.D. Cal. Sep. 12, 2025); *Lee v. R&R Home Care, Inc.*, 2025 WL 2481375, at *4 (E.D. La. Aug. 28, 2025); *Takefman v. Pickleball Club, LLC, 418 So.3d 826, reh'g denied (Sept. 11, 2025)* (Fla. Dist. Ct. App 2025); Order, at 5, *Bischoff v. S.C. Dep't of Educ.*, No. 24-ALJ-30-0362-AP (S.C. Admin. L.Ct. Mar. 6, 2025); Pl.'s Opp'n to Def. Youngblood's Mot. to Dismiss Pursuant to FRCP R. 12(b)(6), at 1, *Jakes v. Youngblood*, No. 2:24-cv-1608 (W.D. Pa. May 27, 2025). Attorneys already cite lack of time as a common reason for not carefully checking AI-generated citations in their own filings. Order to Show Cause, *Dehghani v. Castro*, No. 2:25-cv 00052 (D.N.M. Mar. 11, 2025).

The problem especially impacts pro se litigants, who courts have recognized stand to benefit the most from AI's ability to improve access to justice. *In re Bryant*, No. 25-10147 (Bankr. M.D.N.C. Nov.

19, 2025). Yet they are also the least equipped to detect hallucinated citations. Courts have observed that AI tools "increase [both] pro se litigants' ability to draft voluminous pleadings and increase the odds that such pleadings contain [...] 'hallucinated' legal propositions." *Mitchel v. Stellantis Financial Services, Inc.*, 2025 WL 2676569, at *3 n.6 (E.D. Va. Sep. 18, 2025). Unrepresented litigants lack access to commercial legal databases and are often unaware that AI can fabricate case references. *Al-Hamim v. Star Hearthstone, LLC*, 564 P.3d 1117, 1124 (Colo. App. 2024); *Ligeri v. Amazon.com Services LLC*, 2025 WL 2161497, at *2 (W.D. Wa. July 30, 2025). Courts have responded to these failures with escalating sanctions, finding that earlier, lighter interventions were insufficient to deter the conduct. *Mid Cent. Operating Eng'rs Health & Welfare Fund v. HoosierVac LLC*, 2025 WL 574234, at *3 (S.D.Ind. Feb. 21, 2025). But sanctions are reactive and unevenly applied, and do not address the structural conditions driving the problem (See § 5 for more details).

3. Methods

In § 2, we have shown that hallucination rates are not consistently decreasing, and even if they were, verification burdens would increase nonetheless. Thus, we investigate whether AI can help reduce this burden through automated verification. Because existing datasets do not support evaluation of legal citation verification, we construct the new LEPHANTOMCITE dataset. Based on a taxonomy derived from court cases discussing hallucinated citations, we inject different hallucinated citations in the existing legal brief excerpts.

3.1. Taxonomy of Legal Citation Hallucinations

Non-existent citation. The citation does not correspond to any real case. In Order & Order on Mot. for Leave to File Excess Pages, at 2, *Sims v. Souily-Lefave*, No. 2:24-cv-00831 (D. Nev. Apr. 15, 2025), the pro se plaintiff cited "*Graham v. Nyquist* (1974)", but the citation does not point to any existing case. We generate non-existent citations by altering the reporter volume, reporter abbreviation, or page number to implausible values.

Case name mismatch. The reporter citation and case name refer to two different real cases. In Mem. & Op., at 10-11, *Bevins v. Colgate-Palmolive Co.*, No. 25-576 (E.D. Pa. Apr. 10, 2025), the plaintiff's attorney cited "*Tinch v. Video Indus. Servs., Inc.*, No. 1712 EDA 2018, 2019 WL 1396975, at *3 (Pa. Super. Ct. Mar. 27, 2019)", but the Westlaw citation "2019 WL 1396975" corresponds to *Dolberry v. Jakob*, 2019 WL 1396975 (N.D.N.Y. Mar. 28, 2019) instead. We create mismatches by replacing either the case name or the reporter citation with that of another existing case's.

Incorrect pincite. The citation refers to the correct case, but the cited page number does not support the quoted language or proposition. In *SASC, LLC v. School Supply Connection, Inc., et al.*, 2024 WL 3849424, at *8 (S.D. Ohio Aug. 15, 2024), the attorney did not provide a pincite and was warned to provide pinpoint citations in future filings. The court stated that they will "disregard citations to caselaw that are not supported by pinpoint citations," highlighting the importance of providing correct pincites.

Verbatim misquote. The exact quoted language does not appear in the cited case. In Def.'s Resp. to Pl.'s Mot. for Reconsideration of Order, at 2-4, *Harris v. Take-Two Interactive Software, Inc.*, No. 1:24-cv-01508 (D. Colo. Mar. 6, 2025), the plaintiff cited "*Chambers v. NASCO, Inc.*, 501 U.S. 32 (1991)" in their motion with a quotation that "courts have 'the inherent power to police themselves and to sanction bad-faith litigation conduct." This quote was not found in the cited decision. We generate verbatim misquotes by replacing one or two words in the original quotation with semantically similar

Hallucination type	Modified component	Example
Non-existent citation	reporter citation	Original: 133 S. Ct. 1017 → Hallucinated: 446 Cal. Rptr. 4th 183
Mismatched case name	case name or reporter citation	Original: Cinel v. Connick, 15 F.3d 1338 → Hallucinated: Boone v. Vinson, 15 F.3d 1338
Incorrect pincite	reporter citation, pincite	Original: 830 F.3d at 514 → Hallucinated: 830 F.3d at 511
Verbatim misquote	quote	Original: "[T]he parties' dispute over arbitrability specifically falls within those carve-outs." → Hallucinated: "[T]he parties' dispute over arbitrability specifically resides within those exclusions."
Content misrepresentation	holding	Original: These rules cannot support a claim for retaliatory discharge. → Hallucinated: Any private policies cannot support a claim for retaliatory discharge under Kansas law.

Table 1: **Taxonomy of legal citation hallucinations.** We categorize legal citation hallucinations into five broad categories and show them with examples and relative verification difficulty.

synonyms. This preserves the meaning of the sentence, so there exists no content mismatch (which we define as a separate category below).

Content misrepresentation. A cited case exists but does not support the proposition for which it is cited. In Pl.'s Opp'n to Def. Youngblood's Mot. to Dismiss Pursuant to FRCP R. 12(b)(6), at 1, *Jakes v. Youngblood*, No. 2:24-cv-1608 (W.D. Pa. May 27, 2025), the defendant's attorney wrote in a motion "Moreover, the Pennsylvania Superior Court in Blackwell v. Eskin emphasized that even where statements are embarrassing or upsetting, the Plaintiff must demonstrate their precise defamatory content and origin. [...] 916 A.2d 1123, 1128–29 (Pa. Super. Ct. 2007)." However, the case does not contain any opinion related to identifying defamatory words of the speaker in a complaint. We generate misrepresentation by altering the holdings so that their meanings change and are incorrect. To add realistic LLM-generated content misrepresentations, we additionally include examples from the Large Legal Fiction dataset (Dahl et al., 2024).

These categories are not mutually exclusive: a citation may contain multiple hallucination types (e.g., a mismatched case name and a misquote). However, for dataset construction and evaluation, we only introduce a single type of hallucination to a given citation.

3.2. LePhantomCite Benchmarking Dataset

The LePhantomCite benchmarking dataset contains 1,300 entries drawn from two complementary sources: (1) 1,000 excerpts from real appellate briefs with systematically injected hallucinations, covering all five hallucination types introduced in the previous section; and (2) 300 entries from the LLM-generated central holdings portion of Dahl et al. (2024), which we manually verify and reformat. We combine these sources because Dahl et al. (2024) provides real LLM-generated examples of content misrepresentation. Across the full dataset, there are 4,499 total citation instances, of which

1,107 citation instances contain hallucination. Figure A4 in the Appendix summarizes the count per hallucination type.

Appellate brief excerpts. We collect 245 federal appellate briefs filed in 13 U.S. Courts of Appeals between January 2012 and December 2021, retrieved via the CourtListener API. We restrict to pre-2022 filings to minimize the risk that the source documents themselves contain AI-generated content, and select briefs submitted to Courts of Appeals to ensure high baseline citation quality. Briefs are converted from PDF to plain text using olmOCR (Poznanski et al., 2025), parsed into sentences using a fine-tuned RoBERTa sentence segmentation model (Sanchez, 2019), and then grouped into semantically coherent segments using `Llama-3.3-70B-Instruct` (Grattafiori et al., 2024). This yields 5,648 segments; we subsample 1,000, excluding table-of-contents sections and segments with no case citations.

For each citation in the sampled segments, we use `Qwen3-32B` (Yang et al., 2025) to extract the associated case name, any quotation, and the holding proposition, constrained to exact substrings of the original text to prevent model-introduced errors during extraction. We then inject hallucinations by systematically modifying one or more citation components according to the taxonomy. Hallucinations are introduced into 50% of segments (500 excerpts). We choose this ratio to create a balanced evaluation set, which does not reflect the real hallucination rate observed in generated legal documents.

Non-existent and case name mismatch hallucination types are applied globally: when a citation is replaced with a mismatched case name, all instances of it throughout a brief are altered consistently. This prevents a verification model from detecting the hallucination by cross-referencing correct occurrences elsewhere.

LLM-generated holdings. Content misrepresentation is the hardest hallucination type to inject synthetically. Thus, we supplement the dataset with 300 entries from the central holding task of Dahl et al. (2024), in which LLMs are prompted to state the primary holding of a case given its reporter citation and year. The original hallucination labels in that dataset are based on self-consistency between two LLM outputs rather than ground truth verification. We manually verify all 300 entries against Westlaw, correcting labels where necessary. This process results in 42 confirmed non-hallucinated entries out of 300. See Appendix A1.2 for more details on dataset construction. All holding dataset entries are verified using CourtListener and Westlaw.

3.3. Models and Agent Harnesses

We evaluate five models (agentic and non-agentic). The most successful approach, which we describe in more detail here, is an agentic verification system built on the Bayesian Optimal Experimental Design (BOED) agent framework in Zheng et al. (2026). In BOED, the agent maintains an explicit, language-based belief state that is updated after each action. We adopt this framework because case citation verification is inherently sequential and information-dependent: the agent must extract citations from the brief excerpts, decide on how to gather information and when it has gathered sufficient evidence to make a hallucination determination. BOED's information-directed search loop is well-suited to this structure. We also experiment with the Reflexion agent framework from Shinn et al. (2023), but Reflexion agent struggles to keep track of its current decisions on citation validity when the action sequence grows long. The belief state in BOED solves this issue.

Concretely, the agent's belief state is a language-based record that tracks: (1) all case citations, quotations, and holdings extracted from the input excerpt, and (2) the agent's current assessment: correct, hallucinated, or pending, for each element. The belief state is updated after every action and

serves as working memory across the multi-step verification process, which is necessary because legal excerpts can contain many citations, each requiring independent lookup chains. The agent has access to eight actions, including legal database search and web search (See Appendix Table A4 for list of all actions and corresponding descriptions).

4. Experimental Results

4.1. Experimental Setup

Datasets. We divide the dataset into train and test with a 70-30 split. We run all evaluations on the test set, which contains 390 examples.

Baselines. The agentic framework is model-agnostic. We evaluate it with several open-source and closed-source LLMs, including GPT-5, GPT-OSS-120B (OpenAI et al., 2025), Qwen3-8B (Yang et al., 2025), Qwen3.5-27B, and Gemini-2.5-flash (Comanici et al., 2025). Additionally, we also evaluate non-agentic performance of these models. Experiments using the agentic framework all use $max\_steps = 30$ per episode (one agent run on a single excerpt). As shown in Figure A4 (a), all excerpts contain less than 18 case citations, with a median of 2 citations per excerpt. Some case citations also have associated quotations and holdings, thus we believe 30 steps to be an adequate number to verify all citations within an excerpt. Non-agentic baselines use the same prompt as the agent but without the belief update and action selection sections. We show the full agentic prompt in Appendix A5.

Metrics. We treat the extraction of hallucinated segments from brief excerpts as a retrieval task. Let $G = \{g_1, \ldots, g_m\}$ denote the set of ground-truth hallucinated segments in a brief excerpt and $P = \{p_1, \ldots, p_n\}$ denote the set of segments predicted by the model.

Because predicted spans may not exactly match the annotated spans, we adopt a relaxed matching criterion. Two segments p_1 and g_1 are considered a match if either p_1 is a substring of g_1 or g_1 is a substring of p_1. This still counts predictions that slightly over- or under-span the annotated segment as correct.

Let $\text{match}(p, g)$ be an indicator function that equals 1 if predicted segment $p \in P$ matches a ground-truth segment $g \in G$ under this criterion and 0 otherwise. We compute the number of true positives as the number of predicted segments that match at least one ground-truth segment. We then compute: (1) **Precision** $= \text{TP}/|P|$, the proportion of predicted segments that correspond to a ground-truth hallucinated segment. (2) **Recall** $= \text{TP}/|G|$, the proportion of ground-truth hallucinated segments that are correctly identified by the model. (3) **F1 score** $= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$.

4.2. Results

We show the main results in Table 2, and detailed breakdown by hallucination type in Appendix Table A6. The BOED agentic harness achieves higher recall than non-agentic experiments.

GPT-5 performs overall best, achieving the highest recall on all hallucination types except for incorrect pincites (on which all models perform poorly, see Appendix Table A6). GPT-5 also has the highest average number of steps per episode, averaging at 16.9 steps compared to the lowest, 7.5 steps, from Qwen3-8B. While evaluating agent trajectories, we find that GPT-5 performs the most thorough search, devoting 46.6% of its actions to explore a previously retrieved opinion (action SEARCH_LOCAL_OPINION). It also more thoroughly verifies quotes and stated holdings, consistent with its higher recall on misquote and content misrepresentation categories.

Model	Agentic (max_steps=30)			Non-Agentic		
	Precision (%)	Recall (%)	F1 (%)	Precision (%)	Recall (%)	F1 (%)
Gemini 2.5 Flash	20.2±2.8	65.3±6.4	30.9±3.7	27.8±4.4	45.5±6.2	34.5±4.6
GPT-5	**47.6**±4.7	**82.8**±6.2	**60.5**±4.4	**42.9**±4.7	**63.1**±5.6	**51.1**±4.5
GPT-OSS 120B	24.5±3.4	57.7±5.7	34.4±4.0	17.5±3.2	33.9±5.8	23.1±3.9
Qwen3-8B	15.1±2.4	47.1±5.8	22.9±3.3	15.6±3.0	35.5±5.5	21.6±3.6
Qwen3.5-27B	27.8±3.8	59.8±5.8	38.0±4.3	28.1±5.3	38.7±5.6	32.6±4.9

Table 2: **Agentic and non-agentic model performance.** All models do better using the agentic framework, and GPT-5 achieves the best performance overall.

Appendix Table A6 shows agents' recall performance on each hallucination type. All models (with the exception of Qwen3-8B) find most non-existing cases and case name mismatches, whereas no models can reliably detect wrong pincites. Page number information for many cases is only accessible through Westlaw and LexisNexis, which explains why models fail to recognize these. For content-based hallucinations (verbatim misquotes and content misrepresentation), we find substantive variation in performance: stronger models (GPT-5) can find more than 80% of them, where weaker models only retrieve around 50%.

4.3. Error Analysis

To better understand the potential failure modes of agentic systems, we perform an error analysis on GPT-5 with the BOED harness, the best-performing system, examining all false negatives and false positives in the types non-existent citation, case name mismatch, and verbatim misquote, with the type inferred from the agent's final task beliefs.

False positives on citations absent from CourtListener. The test set contains 129 correct citations that return no lookup result on CourtListener. Any model that flags them produces a false positive because we do not alter these citations. Model behaviour on this subset varies substantially: Qwen3.5 flagged 65.9% of these citations as hallucinated, while GPT-5 flagged only 24.0% (Appendix Table A8). Weaker models appear to treat absence from CourtListener as evidence of hallucination, while stronger models attempt to gather more information from alternative sources before reaching a verdict. Of the 129 citations GPT-5 did not immediately flag, 24 were ultimately verified through alternative sources such as open CourtListener searches or web lookups, while 60 were left pending or unclear due to inconclusive evidence. We show examples of GPT-5 agent leveraging alternative sources to verify citations in Appendix A4.

Exhausted verification budget, sometimes due to redundant actions. 25% of false negatives occur because the agent reached the maximum step limit of 30 before verifying all citations in the excerpt. GPT-5 hit the step limit on 36.7% of test episodes, which suggests that a higher step budget could further improve recall. At the same time, GPT-5 agent does not use its budget efficiently in many episodes. We observe two main types of redundancies. First, when a citation lookup or content search reaches a dead end, the agent often re-issues the same query rather than concluding the citation as hallucinated. The duplicate citation lookup and duplicate opinion search appear in 31.8% and 39.0% of episodes respectively. Second, the agent sometimes re-queries an opinion they have already successfully searched with a slightly different query string. This sometimes destabilizes their belief after a prior search had returned a match. This is observed in 10.5% of episodes (See Appendix Table A9). These redundant searches consume budget without making actual progress.

Reliability failures. Another significant failure mode is the agent's reliability issue (21.4% of

false negatives). A few examples include the following: (1) agent returns output not conforming to the specified format, (2) agent mistakenly returns early without verifying all identified citations, and (3) agent fails to identify all content related to a citation.

Interestingly, agent verification also surfaced over 10 human-made citation typos present in the original pre-LLM briefs (See Appendix Table A1 for details).

5. Discussion and Future Directions

5.1. Information Access Limits Verification Performance

Information access is a binding constraint on verification performance, independent of model capability. Although judicial opinions are public domain, they are not always freely accessible through a single centralized service. For example, the official federal system (PACER) charges per-page fees, and commercial providers routinely bill up to $100 per query (Franklin County Law Library, 2023). While nonprofits like CourtListener have assembled large free databases of opinions, coverage is incomplete and contains no information about whether cited precedent is still good law. Niche cases generated by newer models are more likely to be absent from public repositories, especially if they contain Westlaw-specific identifiers. Additionally, public repositories do not include official reporter pagination, which originates in commercial publisher formatting, making it infeasible to verify the pincites using publicly available sources.

Automated verification tools inherit these same constraints around information accessibility. They directly explain our pincite results and account for a substantial share of false positives from weaker models. Weaker models would interpret absence of a cited case in CourtListener as direct evidence of a hallucination rather than a coverage gap. As shown in § 4.3, the Qwen3.5 model falsely flagged 65.9% of citations not available on CourtListener as hallucinated. Improving automated verification performance will thus require either broader public access to legal databases, or verification systems built on top of commercial platforms. Efforts to broaden free access to federal court records (Schwartz and Albrecht, 2024) are steps in the right direction, but the gap between what is publicly available and what is needed for reliable citation verification remains substantial. Policy efforts to improve equitable access to all case law and legal citations would reduce downstream burdens on the judiciary through improved automated verification methods like those we show here.

5.2. Pro Se Litigants Need Better Support

These information accessibility limitations fall disproportionately on pro se litigants, the group most likely to rely on LLMs for drafting and least likely to have access to commercial legal research databases. The percentage of AI hallucinations in courts attributable to pro se litigants has been steadily rising every year. Courts already postulate that reasons for AI hallucinations in these cases are due to (1) a lack of awareness (See *Al-Hamim v. Star Hearthstone, LLC*, 564 P.3d 1117, 1124, Colo. App. 2024; *Ligeri v. Amazon.com Services LLC*, 2025 WL 2161497, at *2, W.D. Wa. July 30, 2025) and (2) a lack of access to effective verification practices (See *Mitchel v. Stellantis Financial Services, Inc.*, 2025 WL 2676569, at *3 n.6, E.D. Va. Sep. 18, 2025). Courts have responded with sanctions, case dismissals, and, in some jurisdictions, mandatory AI disclosure requirements (Ropes & Gray LLP, 2025). However, these responses are neither designed for nor directed at pro se litigants, and do not foster the potential of AI to increase access to justice for this group.

First, we need effective interventions that educate pro se litigants on how to use AI tools responsibly and equip them with best practices and tools for verifying AI-generated citations. Courts already provide resources for pro se litigants. These should include risks of using AI for pro se litigants, and include detailed AI literacy and citation verification. Second, expanding public, freely-available access to legal data (including all case law) would improve both human and automated citation verification performance for under-resourced filers. For example, Canada's legal information infrastructure is organized around a single free public database unlike the United States. CanLII, a non-profit maintained by the Federation of Law Societies of Canada, provides free access to court judgments from all Canadian courts and all jurisdictions, with generally complete coverage for cases decided after 2001 (Canadian Legal Information Institute, 2025). This structural difference is reflected in how courts reason about pro se verification: one Canadian court noted that verifying whether AI-cited cases exist requires only "a simple search on CanLII," a baseline assumption that US courts cannot reasonably make. *O.K. v Southern Ontario Secondary Schools Association*, 2025 HRTO 2715 at para. 37. Third, automated verification tools could ease verification burdens of courts, with sufficient access to data, but also assist non-lawyers directly to verify citations in documents drafted with AI assistance. Our benchmark provides a step towards developing and auditing such tools.

If courts simply continue to raise sanctions, while not addressing structural limitations around citation verification, AI will ultimately be a false promise for pro se litigation. Facing significant sanctions for potential AI hallucinations and no reliable way to verify citations, pro se litigants may eventually abandon AI altogether. This is in contrast to professional parties who can get access to AI for legal research, and the required tools and practices to ensure responsible usage. Moreover, if access to legal databases becomes a prerequisite for responsible AI use, hiring a professional may ultimately be cheaper and less risky than litigating pro se.

5.3. Future Directions

Despite these information access constraints, we find that agentic systems can already be useful. The agent surfaced over 10 citation typos in pre-LLM briefs that human authors missed. As the volume of AI-generated filings grows, citation verification becomes a more central task. Law clerks, who assist judges by reviewing submitted filings and verifying legal arguments, may take on an expanded role as a result (Katrak and Effoduh, 2026). Our findings suggest that agentic systems can support this work by flagging errors in draft briefs before filing or assisting law clerks in screening submitted documents. However, reliability issues and long-context failures persist and require more research.

The primary aim of this benchmark is to support the development of more reliable and performant citation verification methods. It can further be used to audit existing citation verification tools, which can inform procurement processes of public agencies. However, the semi-synthetic nature of our benchmark means that results may not fully generalize to the distribution of hallucinations in real filings. We therefore encourage future work to complement our dataset with naturally occurring examples. Our benchmark should primarily be treated as a controlled lower bound on verification difficulty.

There are several future directions. In this work, we evaluate the models on short excerpts, but real legal briefs are far longer. The longest brief we collected contains 193 pages, more than 30,000 words, and over 1,000 citations. Agents already struggle with using compute resource efficiently and behaving reliably at the excerpt scale, which are both exacerbated in citation-dense briefs (Rabanser et al., 2026). Hence, future work on legal hallucination detection should focus on alleviating these two issues.

6. Related Work

As LLMs are increasingly applied to the legal domain, numerous works have developed benchmarks and evaluations for hallucinations in legal tasks. These include closed-domain legal summarization (Deroy et al., 2023; Feijo and Moreira, 2021), legal concept explanation (Savelka et al., 2023), and open-domain legal question answering (Dahl et al., 2024; Fan et al., 2026; Hu et al., 2025). Prior work has also proposed taxonomies of legal hallucinations: Dahl et al. (2024) introduces a taxonomy of general legal hallucinations, Magesh et al. (2025) expands it to apply to RAG systems, and Hou et al. (2024) characterize "gaps" between human-written and machine-generated legal analysis, including hallucinations. These benchmarks and taxonomies primarily focus on the correctness of generated legal text. In contrast, our work targets hallucinations in case law citations embedded within legal documents, which is the most common hallucination observed in courts. The closest work to ours evaluates whether LLMs can generate correctly formatted Bluebook citations (Dahl, 2025). This work does not focus on the correctness of citation format, which is a close-domain task, but rather if the citation contains the correct information.

An alternative to reducing hallucinations through model-level improvements is to apply post-hoc safeguards that verify or correct model outputs after generation. Prior work has explored a range of post-hoc hallucination detection and correction methods (Chakraborty et al., 2025). These approaches can be grouped into methods that require access to the original generation process, such as generating multiple responses (Manakul et al., 2023; Mündler et al., 2024; Yehuda et al., 2024) or using specialized prompting strategies (Yu et al., 2024), and methods that operate independently of the generator. The latter category includes training external discriminators to detect hallucinations (Chen et al., 2023b) and systems that query external sources to verify or correct generated summaries (Chen et al., 2023a). Our work aligns with this second class of approaches by treating hallucination detection as a downstream verification problem.

7. Conclusion

Legal citation hallucinations are not a temporary artifact of early LLMs. Across eight ChatGPT generations, we find that legal citation hallucination rates are not consistently decreasing, and that the verification burden on courts is growing along two compounding dimensions: more filings and more citations per filing. These trends will not self-correct as models improve. We thus argue that attention must shift to verification.

We introduce a taxonomy of legal citation hallucinations grounded in real court filings, a dataset of 1,300 brief excerpts with injected hallucinations, and we evaluate an agentic verification system that substantially improves recall over non-agentic baselines. Our results show that there is room for improvement: the best model achieves 60.5% F1, and content misrepresentation — the hallucination type most likely to distort legal outcomes — remains the hardest to detect across all systems. Structural barriers, particularly the absence of official pagination in publicly available legal repositories, further limit what automated verification can currently achieve.

Responsible deployment of AI in legal contexts requires treating verification as a first-class problem. Our taxonomy, dataset, and evaluation framework are a step toward that goal, providing a foundation for building, benchmarking, and auditing the citation checking tools that courts and litigants increasingly need.

References

ABA Law Practice Magazine. Jevons paradox and the legal profession: When efficiency breeds demand. *ABA Law Practice Magazine*, July 2025.

Artificial Lawyer. Jevons paradox + why GenAI will increase legal demand. *Artificial Lawyer*, August 2024.

Canadian Legal Information Institute. What is CanLII. https://www.canlii.org/info/about.html, 2025.

Neeloy Chakraborty, Melkior Ornik, and Katherine Driggs-Campbell. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *ACM Computing Surveys*, 57(7):1–35, 2025.

Cheng-chi Chang. AI, Legal Labor, and the Jevons Paradox. *N.Y.U. Law Review Online*, 101, 2026. forthcoming.

Damien Charlotin. Ai hallucination cases. https://www.damiencharlotin.com/hallucinations/, 2026.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. Purr: Efficiently editing language model hallucinations by denoising language model corruptions, 2023a. URL https://arxiv.org/abs/2305.14908.

Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 245–255, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400701245. doi: 10.1145/3583780.3614905. URL https://doi.org/10.1145/3583780.3614905.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, and Marcel Blistein et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Matthew Dahl. Bye-bye, bluebook? automating legal procedure with large language models, 2025. URL https://arxiv.org/abs/2505.02763.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, January 2024. ISSN 1946-5319. doi: 10.1093/jla/laae003. URL http://dx.doi.org/10.1093/jla/laae003.

Def.'s Resp. to Pl.'s Mot. for Reconsideration of Order, at 2-4, *Harris v. Take-Two Interactive Software, Inc.*, No. 1:24-cv-01508, D. Colo. Mar. 6, 2025. URL https://www.courtlistener.com/docket/68820556/173/harris-v-take-two-interactive-software-inc/.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. How ready are pre-trained abstractive models and llms for legal case judgement summarization?, 2023. URL https://arxiv.org/abs/2306.01248.

Neil Duxbury. *Distinguishing, overruling and the problem of self-reference*, page 111–149. Cambridge University Press, 2008.

Nelson v. Navient Solutions, LLC, 2025 WL 2633962, at *3, S.D. Iowa Sept. 4, 2025.

Al-Hamim v. Star Hearthstone, LLC, 564 P.3d 1117, 1124, Colo. App. 2024.

Fang v. Hechalou US LLC, 2025 WL 3049873, at *3, C.D. Cal. Sep. 12, 2025.

2025 WL 3237890, at *2 *In re Bryant,* No. 25-10147, Bankr. M.D.N.C. Nov. 19, 2025.

Lee v. R&R Home Care, Inc., 2025 WL 2481375, at *4, E.D. La. Aug. 28, 2025.

Ligeri v. Amazon.com Services LLC, 2025 WL 2161497, at *2, W.D. Wa. July 30, 2025.

461 *Mata v. Avianca, Inc.,* 678 F.Supp.3d 443, S.D.N.Y. 2023.

Mid Cent. Operating Eng'rs Health & Welfare Fund v. HoosierVac LLC, 2025 WL 574234, at *3, S.D.Ind. Feb. 21, 2025.

Mitchel v. Stellantis Financial Services, Inc., 2025 WL 2676569, at *3 n.6. , E.D. Va. Sep. 18, 2025.

Noland v. Land of the Free, 114 Cal.App.5th 426, 445, 2025.

O.K. v Southern Ontario Secondary Schools Association, 2025 HRTO 2715 at para. 37. URL https://canlii.ca/t/kg88s.

615 *Park v. Kim,* 91 F.4th 610, 2d Cir. 2024.

Powhatan County Sch. Bd. v. Skinger, 2025 WL 1559593, at *10, E.D.Va. June 2, 2025.

SASC, LLC v. School Supply Connection, Inc., et al., 2024 WL 3849424, at *8, S.D. Ohio Aug. 15, 2024.

Takefman v. Pickleball Club, LLC, 418 So.3d 826, reh'g denied (Sept. 11, 2025), Fla. Dist. Ct. App 2025.

Dongyang Fan, Sebastien Delsad, Nicolas Flammarion, and Maksym Andriushchenko. Halluhard: A hard multi-turn hallucination benchmark, 2026. URL https://arxiv.org/abs/2602.01031.

Diego de Vargas Feijo and Viviane P. Moreira. Improving abstractive summarization of legal rulings through textual entailment. *Artif. Intell. Law*, 31(1):91–113, November 2021. ISSN 0924-8463. doi: 10.1007/s10506-021-09305-4. URL https://doi.org/10.1007/s10506-021-09305-4.

Franklin County Law Library. Lexis & westlaw pricing: Cost-effective electronic legal research, 2023. URL https://fclawlib.libguides.com/costeffectivelegalresearch/pricing. Last updated August 13, 2025. Accessed: 2026-04-10.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, and Chloe Bi et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Abe Hou, William Jurayj, Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. Gaps or hallucinations? scrutinizing machine-generated legal analysis for fine-grained text evaluations. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro, and Gerasimos Spanakis, editors, *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 280–302, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nllp-1.24. URL https://aclanthology.org/2024.nllp-1.24/.

Yinghao Hu, Leilei Gan, Wenyi Xiao, Kun Kuang, and Fei Wu. Fine-tuning large language models for improving factuality in legal question answering. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4410–4427, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.298/.

William Stanley Jevons. *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-Mines*. Macmillan, 1865.

Malcolm Katrak and Jake Okechukwu Effoduh. A few critical people to handle the truth: Ai, hallucinations and the labour of law clerks. In Smita Gupta, Namita Singh Malik, Ardyllis Alves Soares, B. Balamurugan, and Sneha Dhillon, editors, *Artificial Intelligence for Legal System: Jurisprudence in the Digital Age*, pages 121–133. Chapman & Hall / CRC Press (Taylor & Francis), 2026. doi: 10.1201/9781003491903-10.

Grant Lamond. Precedent and Analogy in Legal Reasoning. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2016 edition, 2016.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. Hallucination-free? assessing the reliability of leading ai legal research tools. *Journal of Empirical Legal Studies*, 22(2):216–242, 2025.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Mem. & Op., at 10-11, *Bevins v. Colgate-Palmolive Co.*, No. 25-576, E.D. Pa. Apr. 10, 2025. URL https://www.courtlistener.com/docket/69606630/28/bevins-v-colgate-palmolive-company/.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation, 2024. URL https://arxiv.org/abs/2305.15852.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas

Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Order & Order on Mot. for Leave to File Excess Pages, at 2, *Sims v. Souily-Lefave,* No. 2:24-cv-00831, D. Nev. Apr. 15, 2025. URL https://www.courtlistener.com/docket/68497337/127/sims-v-souily-lefave/.

Order, at 5, *Bischoff v. S.C. Dep't of Educ.*, No. 24-ALJ-30-0362-AP, S.C. Admin. L.Ct. Mar. 6, 2025. URL https://www.polarislab.org/opinion_pdfs/2025-04-10_Bischoff_v_SCDE_order.pdf.

Order to Show Cause, *Dehghani v. Castro,* No. 2:25-cv 00052, D.N.M. Mar. 11, 2025. URL https://www.courtlistener.com/docket/69550312/22/dehghani-v-castro/.

Pl.'s Opp'n to Def. Youngblood's Mot. to Dismiss Pursuant to FRCP R. 12(b)(6), at 1, *Jakes v. Youngblood,* No. 2:24-cv-1608, W.D. Pa. May 27, 2025. URL https://www.courtlistener.com/docket/69412014/45/jakes-v-youngblood/.

Jake Poznanski, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models, 2025. URL https://arxiv.org/abs/2502.18443.

Stephan Rabanser, Sayash Kapoor, Peter Kirgis, Kangheng Liu, Saiteja Utpala, and Arvind Narayanan. Towards a science of ai agent reliability, 2026. URL https://arxiv.org/abs/2602.16666.

Ropes & Gray LLP. Standing orders, local rules, and decisions on the use of AI. https://www.ropesgray.com/en/sites/artificial-intelligence-court-order-tracker, 2025. Last updated June 4, 2025.

George Sanchez. Sentence boundary detection in legal text. In Nikolaos Aletras, Elliott Ash, Leslie Barrett, Daniel Chen, Adam Meyers, Daniel Preotiuc-Pietro, David Rosenberg, and Amanda Stent, editors, *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 31–38, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2204. URL https://aclanthology.org/W19-2204/.

Jaromir Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. Explaining legal concepts with augmented large language models (gpt-4), 2023. URL https://arxiv.org/abs/2306.09525.

David L. Schwartz and Kat M. Albrecht. The SCALES project: Making federal court records free. *Northwestern University Law Review*, 119(1):23–64, 2024. URL https://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=1571&context=nulr.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

Brandon Smith and Anton Troynikov. Evaluating chunking strategies for retrieval. Technical report, Chroma, July 2024. URL https://research.trychroma.com/evaluating-chunking.

The Bluebook Editors. *The Bluebook: A Uniform System of Citation*. Harvard Law Review Association, 22nd edition, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9333–9347, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.506. URL https://aclanthology.org/2024.acl-long.506/.

Xiaodong Yu, Hao Cheng, Xiaodong Liu, Dan Roth, and Jianfeng Gao. ReEval: Automatic hallucination evaluation for retrieval-augmented large language models via transferable adversarial attacks. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1333–1351, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.85. URL https://aclanthology.org/2024.findings-naacl.85/.

Lucia Zheng, Zeyu Shen, Boyi Wei, Kylie Zhang, Max Gonzalez Saez-Diez, Nimra Nadeem, Kincaid MacDonald, Liam H Fowl, Zirui Cheng, Thomas L. Griffiths, Daniel E. Ho, Christopher D Manning, Dilip Arumugam, and Peter Henderson. Learning to explore through information-directed bayesian optimal experimental design, 2026. *(unpublished manuscript) (on file with author).*

A1. Dataset Construction

We describe the dataset curation process in more detail in this section.

A1.1. Bluebook Legal Citation Format

Verifying a legal citation is not a single check but a structured set of checks over each component that the citation encodes. Legal citations in U.S. court filings follow the Bluebook, a standardized citation system used across courts and legal practice (The Bluebook Editors, 2025). A standard case citation consists of five components: (1) a *case name*, identifying the parties, (2) a *reporter citation*, specifying the volume, reporter, and first page of the decision, (3) an optional *pincite*, indicating the specific page(s) supporting a proposition or quotation, (4) a *parenthetical*, providing the court and year of decision, and (5) other parenthetical information, and subsequent history of the case, if any. For example, in a citation of the form *Aves v. Shah*, 997 F.2d 762, 767 (10th Cir. 1993), *Aves v. Shah* is the case name, 997 F.2d 762 is the reporter citation, 767 is the pincite of a particular page, 10th Cir. is the court, and 1993 is the date of decision. All components should be consistent to form a correct citation.

A1.2. Legal Hallucination Dataset

We collect a dataset of legal briefs that were filed in courts. We use the CourtListener search API to find federal appellate briefs that were filed between 2012-01-01 and 2021-12-31. We downloaded 323 documents in total in PDF format. We then convert all PDFs to plain text using olmOCR (Poznanski et al., 2025).

To construct the hallucination dataset, we proceed in several steps. First, we use the CourtListener citation lookup API to extract all case citations from the collected briefs, yielding a total of 27,949 distinct legal citations. 2,854 citations of these are not found on CourtListener. Most of these citations are Westlaw or Lexis citations. We do not alter any citations that are not found on CourtListener because these citations are already challenging to verify given the lack of access. Figure A1 shows the number of citations per brief distribution. Most briefs have between 0 to 200 citations. We discard briefs with over 200 citations, leaving 279 briefs. After injecting hallucinations and breaking briefs into paragraphs, we subsample again to form the final dataset that contains 1,000 excerpts. The excerpts in the final dataset come from 245 briefs.

For each citation, we extract a local context from the brief. Next, we use the `Qwen3-32B` (Yang et al., 2025) models to identify the associated case name and, when present, the corresponding quotation as well as the holding each citation is used to support based on the surrounding context. To decrease the risk of hallucination, we use LLMs solely to identify and retrieve structured citation components from the original briefs. We also ensure that all the retrieved case names, quotations, and holdings appear verbatim in the original text.

Finally, we inject hallucinated citations by systematically modifying one or more components of the original citation according to the taxonomy in Section 3.1. To generate misquotes, we use `Qwen3-32B` to alter one or two words in the quotes. We discard any generations that are significantly different from the original quotes. To generate content misrepresentation, we use `Qwen3-32B` to alter the holding to be factually incorrect while remaining plausible given the surrounding context. The model prompts used to extract and alter quotes and holdings are in Appendix A5. For each brief, we randomly sample citations for hallucination injection, and use weighted random sampling to select

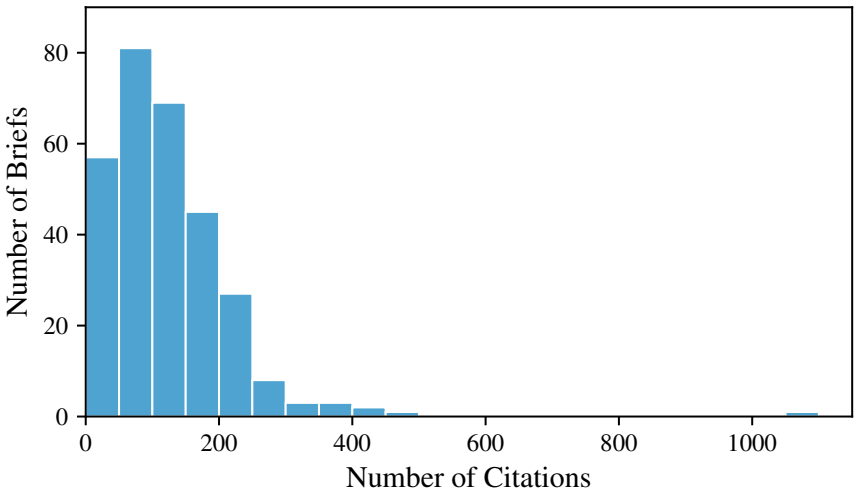


Fig. A1: **Number of citations per brief.** Most briefs have 0-200 citations.

the hallucination type to inject. Not all case citations are eligible for all types of hallucination. For example, only citations that are associated with quotes are eligible for misquotes, while all citations are eligible for non-existent citation. To ensure balanced distribution across all types, we use weighted sampling to choose hallucination type. Additionally, we apply non-existent citations and case name mismatches globally across all instances of the same case citation. We apply these globally to ensure that the verification model cannot infer hallucination based on correct citations of the same case that appear in other parts of the brief. For example, one instance of the citation *Cinel v. Connick, 15 F.3d 1338* is altered to *Boone v. Vinson, 15 F.3d 1338*, but if the correct citation *Cinel v. Connick, 15 F.3d 1338* appears in another sentence, the model might be able to infer that the citation contains some inaccuracies solely based on the inconsistency.

Then, we parse full briefs into segments. We use segments instead of full briefs as unit for verification to standardize the process. The briefs have a wide range of word counts (See Figure A2). We apply LLM-based semantic segmentation (Smith and Troynikov, 2024): We first fine-tune a RoBERTa model to do legal sentence segmentation (Sanchez, 2019). We then use this model to segment briefs into sentences. Afterwards, we use a Llama3.3-70B model to group sentences into semantically coherent paragraphs, resulting in 5,648 segments. We subsample 1,000 from all resulting segments. When sampling, we exclude all segments that contain Table of Content because they only contain lists of case citations without any context, making them less challenging. We also exclude contact information of attorneys and only sample segments that contain at least one citation.

We only use the central holding task portion of the Legal Hallucinations dataset from Dahl et al. (2024) because it is the closest to our task. However, it is a QA dataset, so we need to re-format it to be aligned with our dataset input. The original dataset is constructed by prompting LLMs to output the primary holding of a case given its reporter citation and year. It contains 1494 entries based on 300 unique cases and has outputs from 4 LLMs. Each entry has a binary *hallucination* label, a *correctness_score* ranging from 0-100, additionally with -99 indicating invalid response, and two LLM responses *llm_output_1* and *llm_output_2*.

Each case has multiple entries corresponding to responses from the different LLMs. We subsample from the whole dataset so that each case has one entry, resulting in 300 entries. When sub-

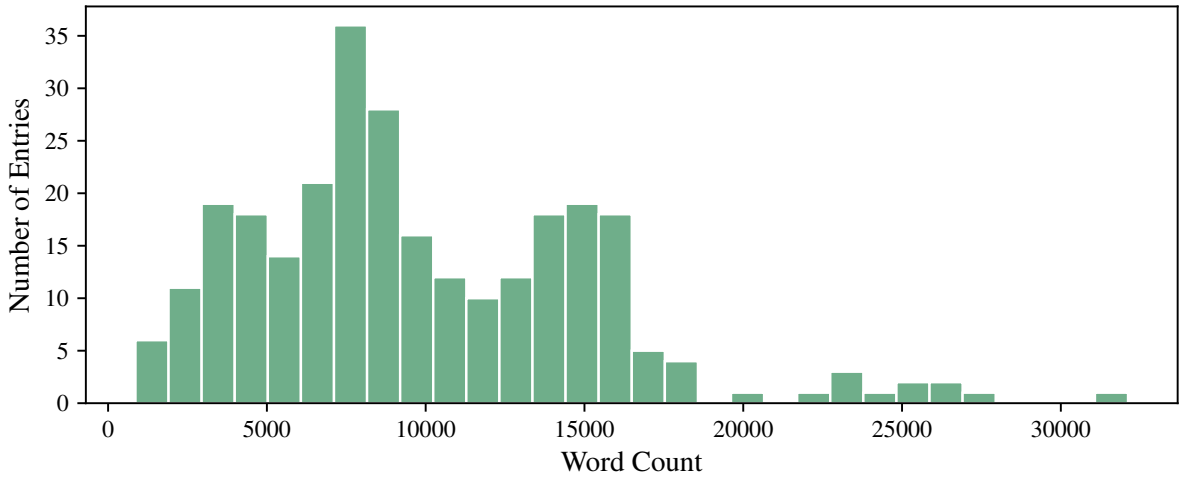


Fig. A2: **Word count per brief.** The variance in word count of briefs is very high.

sampling, we prioritize sampling entries with no hallucination to result in a more balanced final dataset. There are 115 non-hallucinated responses and 185 hallucinated responses in the subsampled dataset according to the original *hallucination* label. We transform the dataset format to align with our dataset's. We construct each entry:

> **Hallucinations:** [<*llm_output_1*>] if *hallucination* is True, otherwise [].
> **Text:** "<*llm_output_1*>. See <*case name*>, <*citation*> (<*year*>)."

llm_output_1, *citation*, and *year* are fields in the Legal Hallucinations dataset, and *case name* is obtained from a CourtListener lookup of the citation.

However, the *hallucination* labels are not based on whether the LLM outputs are correct but whether the two outputs are consistent with each other. This leads to a lower bound measure of hallucination rate, but for our dataset we need accurate hallucination labels, so we manually verify all labels.

We verify all 300 entries using Westlaw. We leave the text as is if it is consistent with its *hallucination* label. If there is inconsistency in the stated case year and the Westlaw opinion year, we use the year from Westlaw.

Combining the two parts of the dataset, there are 1,300 entries in total.

Figure A3 shows the distribution of the number of times each case in the dataset has been cited. We see that both hallucinated and non-hallucinated cases follow the same citation distribution. The model thus cannot solely rely on knowledge of landmark cases to check for hallucinations.

Example datapoint. Each entry in the dataset consists of a short excerpt from a legal brief and a list of ground truth hallucinated segments with the hallucination type label. The task is to extract all hallucinated segments from the excerpt. Below is an illustrative example involving incorrect pincite and content misrepresentation:

> **Hallucinations:** {"370 F.3d 1223, 1224": "incorrect_pincite", "US Parole Commission members are 1983 persons when they act pursuant to the Eighth Amendment's prohibition on cruel and unusual punishment.": "Content misrepresentation"}

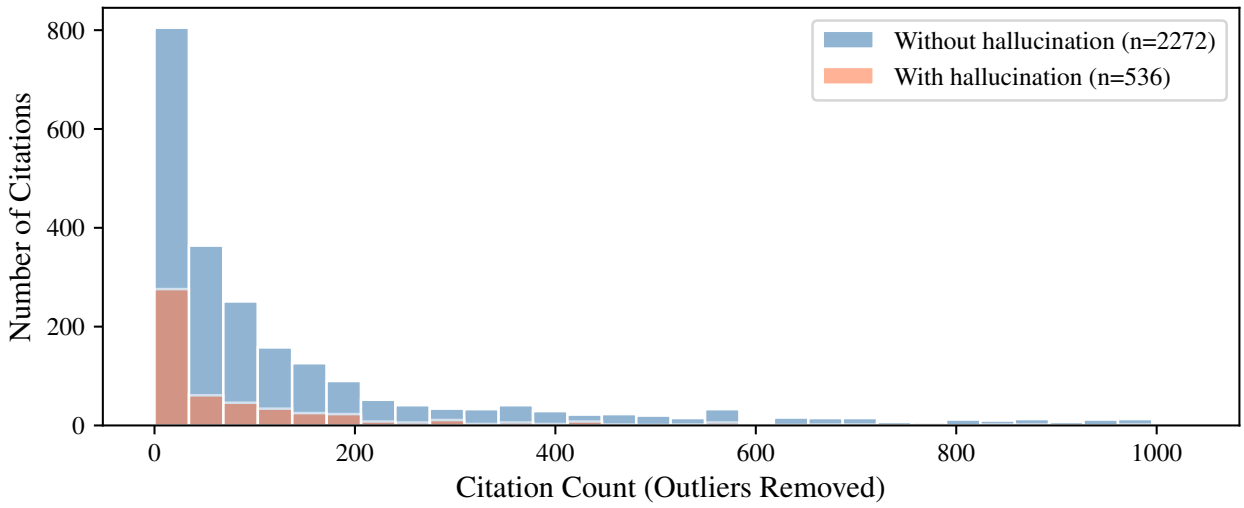


Fig. A3: **Number of citations cases have.** The hallucinated (red) and non-hallucinated (blue) cases have the same citation distributions. We remove the outliers when plotting because a few cases are cited many times that they would skew the visualization.

> **Text:** "[...] but the Commissioners exercise District powers when they handle D.C. parolees. See, e.g., Fletcher v. Dist. of Columbia, 370 F.3d 1223, 1224 judgment vacated on reh'g on other grnds, 391 F.3d 250 (D.C. Cir. 2004) US Parole Commission members are 1983 persons when they act pursuant to the Eighth Amendment's prohibition on cruel and unusual punishment. [...]"

In this example, 370 F.3d 1223, 1224 is hallucinated because the pincite is incorrect, and the holding is misrepresented because it should be "pursuant to D.C. Revitalization Act" instead of "pursuant to the Eighth Amendment's prohibition on cruel and unusual punishment".

A1.3. Dataset Label Update

We create the ground truth hallucination segments of each dataset entry under the assumption that the original briefs do not contain any mistakes that are similar to model hallucinations. We select briefs from 13 U.S. Courts of Appeals to ensure quality. However, humans make mistakes as well. We examined all false positive outputs from GPT-5 agent of the types non-existent citation, case name mismatch, and misquote (hallucination types are inferred from the final task beliefs). We found some typos in the original briefs, as well as some minor stylistic choices that are ambiguous in terms of whether they should be counted as mistakes or not. We update the ground truth hallucination labels to include the clear typos (See Table A1). We create another category of optional ground-truth labels to include stylistic choices (e.g., "fact checking" vs. "factchecking") and other instances where we cannot locate the case on Westlaw but are also not confident that the case does not exist (see Table A2). An optional ground truth segment means that a prediction matching these spans counts as a true positive, but failing to predict them does not count as a false negative.

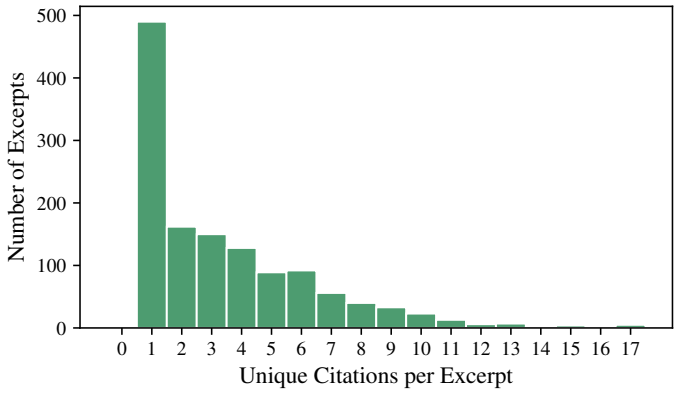
File	Type	Optional Span
`Allina Health v Sebelius`	Case name mismatch	Heartland Regional Med.Ctr. v. Sebelius

Kimberlin v Frey	Verbatim misquote	a loan servicer will become a debt collector under 1692a(6)(F)(iii) if the debt was in default or treated as such when it was acquired.
Clear with Computers v Altec	Verbatim misquote	meaningful limitations
Estate of Bonilla v City of York	Verbatim misquote	raise an inference
Evolutionary Intelligence v Sprint	Content misrepresentation	involves a specific system for modifying data that has equally concrete and valuable effects in [its field].
Evolutionary Intelligence v Sprint	Incorrect pincite	Diehr, 101 S.Ct. at 1048
Fortres Grand v Warner Bros.	Verbatim misquote	unfair competition category
Carrera v Bayer Corp	Non-existent citation	135 S. Ct. 940 (2013)
Carrera v Bayer Corp	Verbatim misquote	best practicable notice
Gonzalez v New Life Ventures	Verbatim misquote	It is obvious that the subject matter described in 101 is expansive. As the Supreme Court has observed, the subject-matter provisions of the patent law have been cast in broad terms to fulfill the constitutional and statutory goal of promoting the Progress of Science and the useful Arts.
Hockeyline v Stats LLC	Verbatim misquote	this subsidiary fact finding must be reviewed for clear error on appeal.
Nemphos v Nestle Waters N. Am.	Verbatim misquote	"Pure Water Perfect Taste"
Rentmeester v Nike	Verbatim misquote	a court must filter out and disregard the non-protectable elements in making its substantial similarity determination.
Jones v Citimortgage	Verbatim misquote	a weighing of evidence on both sides,
Jones v Citimortgage	Verbatim misquote	mechanically determined through relatively rigid legal rules.
Westberg v FDIC	Verbatim misquote	As a practical matter of statutory construction, … we proceed on the assumption that Congress intended the 'claims' barred by 1821(d)(13)(D) to parallel those contemplated under FIRREA's administrative claims process laid out in the greater part of 1821(d).
Belleau v Wall	Non-existent citation	Ellis v. State, 353 S.E.2 19 (Ga. 1987)
Par Pharmaceutical v TWI	Verbatim misquote	[u]se of prophetic examples does not automatically make a patent non-enabling.
Smith v Aegon Pension Plan	Verbatim misquote	are generally free under ERISA, for any reason, at any time, to adopt, modify, or terminate welfare plans
United States v Velasquez	Content misrepresentation	The prosecutor's comment during closing argument was more prejudicial. It was a direct comment on defendant's failure to testify, and it occurred only one day before the jury began deliberations.
United States v Flanders	Verbatim misquote	In general, a sentence within the limits imposed by statute is neither excessive nor cruel and unusual.
W. Reserve Life v DK LLC	Non-existent citation	715 F. Supp. 270 (D.R.I. 2010)

Table A2: **Optional ground-truth segments.** A prediction matching these spans counts as a true positive, but failing to predict them does not count as a false negative.

A1.4. Controlled Experiment LLM-Generated Legal Documents

We generate prompts in this experiment by collecting legal filings with AI-generated hallucinations and reverse-engineering the prompts that were used to generate them. We show example prompts in Table A3.



Hallucination type	Count
Non-existent citation	158
Mismatched case name	189
Incorrect pincite	177
Verbatim misquote	167
Content misrepresentation	416
Total	1,107

Fig. A4: **Legal citation verification dataset summary.** Distribution of number of unique citations in each brief excerpt (left) and the number hallucinated citations belonging to each hallucination type (right).

File	Original	Correct
`Abbey v United States`	Gieg v. DRR, Inc.	Gieg v. DDR, Inc.
`Abbey v United States`	216 F.3d 659	276 F.3d 659
`Crystallex Int'l v Petroleos`	Shenago, Inc.	Shenango, Inc.
`Delaware Riverkeeper v FERC`	egregious or arbitrary	[egregious] or [arbitrary]
`Gametek v Zynga`	'so theoretical and broad'	'so abstract and sweeping'
`Gametek v Zynga`	Benson, 409 U.S. at 67	Benson, 409 U.S. at 68
`In re Blaine Keith`	587 U.S. 460	567 U.S. 460
`In re Blaine Keith`	Fuentes v. Shervin	Fuentes v. Shevin
`Intercollegiate Broadcasting v CRB`	vesting authority	vesting of authority
`Jones v Citimortgage`	Bonano v. Thomas	Bonanno v. Thomas
`Loveridge v Hall`	"jurisdiction over all"	"jurisdiction of all"
`Nozzi v HACLA`	avoiding an abrupt change	an abrupt and unexpected change
`United States v Esquenazi`	Id. at 815-16	Id. at 809
`Westberg v FDIC`	624 F.3d at 1144	642 F.3d at 1144
`Witham v New York State`	Roadway Express v. Pipe	Roadway Express v. Piper
`Witham v New York State`	Sims v. Aherns	Sims v. Ahrens
`Zzyym v Pompeo`	Plyer v. Doe	Plyler v. Doe

Table A1: **Typographic errors found in original briefs** We found these minor errors in the original briefs when examining the false positives in agent outputs. We update the ground truth labels to include them.

	Prompt Text
Prompt example 1	Generate an Objection to Report and Recommendation in the Denial of In Forma Pauperis Application. Cite relevant cases to support the objection. The following facts are given: The plaintiff, Bonnie Nicole Luther, is objecting to a Magistrate Judge's recommendation to deny her application to proceed in forma pauperis (without prepayment of fees or costs). She argues that she cannot afford the filing fees and should be granted IFP status. Her reasons for objecting include: 1. Lack of Financial Capacity: She claims to have no personal income or public benefits, and although she has a joint bank account with her husband, the funds are exclusively for essential household expenses for their family of five, and she lacks control over discretionary finances. She contends that mere joint account access does not equate to disposable income. 2. Joint Property Status: While she is a joint owner of her home, she states she receives no financial benefit from it, has no accessible equity, and it is a primary residence, not a liquid asset. 3. Good Faith Submission: She asserts that she completed her IFP applications in good faith, responding based on her actual knowledge and access. 4. Importance of Legal Claims: The underlying lawsuit involves serious allegations of civil rights and constitutional violations against state actors under federal law, and denying IFP status would prevent her from pursuing these claims due to poverty. She emphasizes that the in forma pauperis statute is meant to ensure equal access to the judicial system for indigent persons, especially in cases involving governmental abuse or constitutional harm.
Prompt example 2	Generate an Objection to the Magistrate Judge's Report and Recommendation Denying In Forma Pauperis Application. Cite relevant cases to support the objection. The following facts are given: The plaintiff, Bonnie Nicole Luther, is objecting to a Magistrate Judge's recommendation to deny her application to proceed in forma pauperis. She asserts the denial was incorrect and offers the following reasons: 1. Irregular and Insufficient Income: She works part-time as a substitute teacher but earns well below the federal poverty line. Her hours are unpredictable, and during several months she earns no income at all. She argues this erratic income should not disqualify her from IFP status, especially as she has no savings and is behind on several utility and medical bills. 2. Debt and Financial Obligations: Bonnie carries significant unsecured debt, including past-due credit card balances and a defaulted student loan. Additionally, she is the primary caregiver for an elderly parent with no income of their own, which imposes significant caregiving and financial burdens. 3. Minimal Liquid Assets: While she does have a small individual checking account, the balance rarely exceeds $50 and is depleted monthly by basic living expenses. She owns no property, has no investments, and does not own a car. 4. Good Faith and Transparency: She emphasizes that she disclosed all requested financial information and attachments with full transparency, and no part of her application was intended to mislead or omit material facts. 5. Meritorious and Time-Sensitive Legal Claims: Her lawsuit raises important claims under the Americans with Disabilities Act and Section 1983 for denial of reasonable accommodations and violations of her First and Fourteenth Amendment rights. She contends that denying her access to court based solely on poverty is contrary to the purpose of § 1915 and would irreparably harm her ability to seek redress for ongoing constitutional violations.

Table A3: Example prompts used to generate legal documents for LLM legal citation hallucination rate estimation.

A2. Additional Agentic Framework Details

A2.1. Agent Actions

Type	Action	Description
Lookup	`COURTLISTENER_CITATION_LOOKUP`	Calls the CourtListener lookup API endpoint given a string of text containing a case citation.
	`OPEN_COURTLISTENER_SEARCH`	Performs an open search on CourtListener using keywords.
	`OPEN_WEB_SEARCH`	Searches the web with SerpAPI for information.
Search	`ACCESS_COURTLISTENER_OPINION`	Calls the CourtListener opinion API endpoint to access an opinion.
	`SEARCH_LOCAL_OPINION`	Searches text in the local opinion.
Reading	`READ_DOCUMENT`	Reads selected lines from the chosen opinion.
	`EDIT_SCRATCHPAD`	Takes notes when reading documents.
Reasoning	`THINK`	Performs internal reasoning.

Table A4: **Agent action types and their descriptions.** The agent has access to eight different actions grouped into four types.

A2.2. Model Specification

Model	Access	Params	Temp.	Max Tokens	GPUs	Quant.
Gemini 2.5 Flash	API (Gemini)	—	0.8	8,192	–	–
GPT-5	API (OpenAI)	—	0.8	10,000	–	–
GPT-OSS-120B	Local (vLLM)	120B	0.8	8,192	2	MXFP4
Qwen3.5-27B (FP8)	Local (vLLM)	27B	0.8	8,192	1	FP8
Qwen3-8B	Local (vLLM)	8B	0.8	8,192	1	MXFP4

Table A5: **Model configurations used in experiments.**

Hallucination Type	Gemini 2.5 Flash	GPT-5	GPT-OSS 120B	Qwen3-8B	Qwen3.5-27B
Non-Existent Citation	90.3±12.1	**100.0**±0.0	90.3±10.3	58.1±18.9	96.8±5.6
Case Name Mismatch	85.3±8.4	**97.1**±3.8	83.8±10.5	52.9±11.4	94.1±5.2
Incorrect Pincite	16.4±12.0	18.2±10.7	21.8±11.4	16.4±11.2	**25.5**±12.6
Verbatim Misquote	67.4±14.8	**82.6**±11.1	39.1±15.0	32.6±14.6	32.6±13.7
Content Misrepresentation	55.7±8.6	**84.0**±6.2	53.4±8.7	54.2±8.9	49.6±8.7

Table A6: **Agentic model recall by hallucination types.** All models perform the worst on content misrepresentation and incorrect pincite tasks. We only show the recall in this table because the models do not lable the hallucination type when outputting hallucinated segments, thus making per type precision inaccurate.

Action Type	Gemini 2.5 Flash	GPT-5	GPT-OSS-120b	Qwen3-8B	Qwen3.5-27B
ACCESS_COURTLISTENER_OPINION	21.8%	15.2%	17.0%	16.1%	23.3%
COURTLISTENER_CITATION_LOOKUP	30.0%	21.0%	33.3%	49.6%	28.9%
EDIT_SCRATCHPAD	0.2%	0.0%	0.0%	0.0%	0.3%
OPEN_COURTLISTENER_SEARCH	7.6%	6.9%	10.0%	9.8%	6.1%
OPEN_WEB_SEARCH	2.9%	8.7%	6.2%	3.6%	1.6%
READ_DOCUMENT	2.7%	1.6%	3.3%	3.5%	3.8%
SEARCH_LOCAL_OPINION	33.6%	46.6%	30.1%	16.4%	35.3%
THINK	1.2%	0.0%	0.1%	1.0%	0.8%
Average # of Tool Calls	10.9	16.9	17.7	7.5	10.2

Table A7: **Action type frequency by model.** GPT-OSS-120b takes the highest average number of steps on each data point. This is likely due to it having the highest percentage of duplicate citation lookups (See Table A9).

A3. Additional Experimental Results

A3.1. Model Performance by Hallucination Type

Table A6 shows the agents' performance on different hallucination types. Most models achieve high recall on non-existent citations and case name mismatches

A3.2. Agent Tool Call Analysis

Table A7 shows the distribution of action types taken by each model during evaluation. Across all models, the three most frequently used actions are COURTLISTENER_CITATION_LOOKUP, ACCESS_COURTLISTENER_OPINION, and SEARCH_LOCAL_OPINION, reflecting that legal database access is the core bottleneck in citation verification. The primary difference between stronger and weaker models is not the number of lookups performed, but the actions taken after a successful lookup. GPT-5 devotes 46.6% of its actions to SEARCH_LOCAL_OPINION, the highest of any model, while Qwen3-8B, the weakest performer, spends only 16.4% on this action and instead concentrates over half its steps on direct citation lookups. This pattern suggests that after locating a citation, stronger models continue into the opinion text to verify quotes and holdings thoroughly. This is consistent with GPT-5's stronger recall on misquote and content misrepresentation categories, both of which require reading the underlying opinion rather than merely confirming a citation exists. GPT-5 also uses OPEN_WEB_SEARCH more than any other model (8.6% of actions), likely as a fallback strategy for citations absent from CourtListener. This behavior is examined further in the error analysis below.

A3.3. Error Analysis

Handling citations not on CourtListener. In the test set, there are 129 citations not found on CourtListener. As described in Section 3.2, these citations were not altered during hallucination injection and are therefore all correct. Any model that flags them as hallucinated produces a false positive. Table A8 shows the false positive rate on these citations across models. The variation is substantial: Qwen3.5 flagged 65.9% of them as hallucinated, while GPT-5 flagged only 24.0%. This gap reflects a meaningful difference in verification strategy. Weaker models appear to treat absence from CourtListener as evidence of hallucination, which is a brittle heuristic given that roughly 10% of real citations in our dataset are Westlaw or Lexis citations that CourtListener cannot resolve.

GPT-5's handling of these cases is more sophisticated. Of the 129 citations missing from CourtListener, it correctly withheld a hallucination judgment on 98. Of these, 41 were ultimately verified through alternative sources: open CourtListener searches occasionally returned case dockets that confirmed a citation's existence even without resolving the Westlaw or Lexis identifier, and in some cases GPT-5 located opinions that cited the same reference. A further 40 citations were left in a pending state, meaning that the agent searched but could not reach a conclusive determination. The remaining cases were never checked, typically due to unusual citation formats or non-standard sources such as FERC citations.

This pattern illustrates an important trait of effective verification agents: the ability to recognize the limits of a single database, seek corroborating evidence across multiple sources, and calibrate confidence accordingly rather than defaulting to a false positive.

Duplicate actions that do not generate new information. In many episodes, the agents take near-duplicate actions that do not generate new information. We classify duplicate actions into three categories:

1. Duplicate citation lookup: COURTLISTENER_CITATION_LOOKUP is called with the same citation string (e.g. "708 F.3d 704") more than once in the same episode.
2. Duplicate opinion search: SEARCH_LOCAL_OPINION is called with the exact same (opinion_id, query) pair more than once.
3. Re-search after hit: SEARCH_LOCAL_OPINION returns found=True for an opinion, but the model later calls SEARCH_LOCAL_OPINION on the same opinion X again with a different query that is $> 50\%$ similar to the one that already succeeded.

Duplicate citation lookup and duplicate opinion search mainly happen when the agent reaches a deadend with its queries. However, instead of recognizing that the citation or the information it is verifying is hallucinated, it attempts to restart to find different information. Re-search after hit type of duplicate happens when the agent tries to confirm a quote or a holding repeatedly. When a follow-up search does not return a result due to changed wording, the agent's belief on the citation becomes unstable even if previous search returned a match and should have confirmed its correctness.

Model	# Flagged	False Positive Rate (%)
Gemini 2.5 Flash	76	58.9
GPT-5	31	**24.0**
GPT-OSS 120B	56	43.4
Qwen3-8B	67	51.9
Qwen3.5-27B	85	65.9

Table A8: **False positive rate on citations not found on CourtListener.** GPT-5 has the lowest false positive rate on these citations. When a citation is not found on CourtListener, it usually attempts to find it from other sources.

Model	Duplicate Citation Lookup (%)	Duplicate Opinion Search (%)	Re-search After Hit (%)	All (%)
Gemini 2.5 Flash	21.8	11.3	2.8	25.1
GPT-5	31.8	39.0	10.5	53.3
GPT-OSS 120B	55.6	35.1	12.3	62.8
Qwen3-8B	35.4	**3.6**	**0.3**	36.2
Qwen3.5-27B	**12.1**	8.5	1.8	**18.7**

Table A9: **Percentage of episodes containing each type of redundant tool call.** Many models perform duplicate calls that do not generate new information.

A4. Case Studies of Agent Trajectory

In this section, we show a few agent trajectories where the agent is able to leverage information on the internet to verify citations and their contents.

Verifying citation validity using external sources. In the first example, the agent looked up a citation, "1996 WL 451054", that is not available on CourtListener. Instead of concluding that the citation is hallucinated after getting no result from CourtListener, the agent searched up the citation on the Internet. It found a result from a Michigan Law School clearinghouse PDF that states "1996 WL 451054. Only the Westlaw citation is currently available. United States District Court, District of Columbia. Joy EVANS, et al...". The returned snippet confirmed both the existence of the citation and the case name consistency. With this information, the agent confirmed the validity of the citation, while the validity of the holding it supports remained pending.

GPT-5 agent trajectory 1

TRAJECTORY: david-harvey-v-leon-mohammed_1289365_part34.pdf
Model: gpt-5
Total Steps: 30
Brief excerpt contains: "Evans, 1996 WL 451054"

==========

STEP 16 — COURTLISTENER_CITATION_LOOKUP Cite: 1996 WL 451054
Result: NOT FOUND (404) — "Citation not found: '1996 WL 451054'"

TASK BELIEFS (entering step 17):
• Evans, 1996 WL 451054, at *1–*2 — pending (CL lookup returned 404; may be Westlaw-only or outside CourtListener's coverage; further verification via alternative sources needed)

STEP 17 — OPEN_WEB_SEARCH
Query: 1996 WL 451054
Results:
[1] "Order Adopting Proposed Findings of Fact of Special Master"
URL: https://clearinghouse-umich-production.s3.amazonaws.com/ media/doc/63437.pdf
Snippet: "1996 WL 451054. Only the Westlaw citation is currently available. United States District Court, District of Columbia. Joy EVANS, et al..."
[2] govinfo.gov / D. Del. court doc
[3] ca8.uscourts.gov

TASK BELIEFS (after step 17 observation):
• Evans, 1996 WL 451054, at *1–*2 — VERIFIED AS CORRECT for citation existence and case identity within Evans litigation. Evidence: Civil Rights Litigation Clearinghouse PDF explicitly references "1996 WL 451054" and "Joy Evans, et al." CourtListener coverage is unavailable. Content/pincite still pending.

Verifying citation content using external sources. The second example illustrates a more complex verification path: instead of searching for only the citation on the internet, the agent had to search the holding as well to find a relevant and useful source. The agent first searched the citation and the name on CourtListener and the Internet, but the results did not confirm the citation. The Internet search results confirmed that the case name and docket ID are consistent (See Step 7 and the task belief after it), but it did not confirm the Westlaw citation from the excerpt. Then, the agent tried to

search the case name and the Westlaw citation together on CourtListener and the Internet, but it did not get any new information.

The key step is Step 20, where the agent searched for the content of the case stated in the brief excerpt on the Internet. This time it found the actual case document hosted govinfo.gov as a PACER-sourced PDF that directly confirmed the content of the case. The date of the document also aligned with the Westlaw citation, so the agent concluded that the Westlaw citation is likely correct even though it lacked concrete evidence. The pincite still could not be verified using freely accessible information.

GPT-5 agent trajectory 2

TRAJECTORY: charlotte-weatherford-v-nevada-rural-housing-authority_664577_part18.pdf
Model: gpt-5
Total Steps: 30
Brief excerpt contains: "See also Alamar Ranch, LLC v. County of Boise, CV 09-04-S-BLW, 2010 WL 1727844, *5 (D. Idaho Apr. 27, 2010) (a reasonable accommodation must be for the benefit of the disabled, and not merely the personal gain of the requester)"

==========

STEP 6 — OPEN_COURTLISTENER_SEARCH Query: Alamar Ranch LLC v. County of Boise
Result: No matching opinions found on CourtListener.

TASK BELIEFS (after step 6):
5) Alamar Ranch, 2010 WL 1727844 — PENDING. CL search returned no direct match; related litigation appears in a creditor reference (In re Boise County). Specific WL-cited opinion unverified. Next step: search by docket CV 09-04-S-BLW / 1:09-cv-00004-BLW.

STEP 7 — OPEN_WEB_SEARCH Query: 2010 WL 1727844 Alamar Ranch LLC v. County of Boise Results:
[1] "Alamar Ranch v. County of Boise" URL: https://www.jenner.com/.../Alamar_Ranch_v._County_of_Boise.pdf Snippet: "For the reasons expressed below, the Court will grant in part the motion for return of privileged documents and deny the motion to strike."
[2] "Alamar Ranch, LLC v. County of Boise, 1:09-cv-00004" URL: https://app.midpage.ai/case/alamar-ranch-llc-v-county-... Snippet: "D. Idaho. Nov 1, 2009. Read the full case. Opinion summary not found."

TASK BELIEFS (after step 7):
5) Alamar Ranch, 2010 WL 1727844 — PENDING. New info: related D. Idaho docket confirmed (1:09-cv-00004); case exists but specific Apr. 27, 2010 WL-cited opinion remains unverified. Docket/case name matches.
6) Alamar Ranch holding — PENDING

STEP 12 — COURTLISTENER_CITATION_LOOKUP Cite: 2010 WL 1727844
Result: NOT FOUND (404) — "Citation not found: '2010 WL 1727844'"

TASK BELIEFS (after step 12):
5) Alamar Ranch, 2010 WL 1727844 — PENDING. CL returned 404; increasing likelihood of WL-only or unindexed order; docket may also be styled "v. Boise County" with variation "CV-09-004-S-BLW."
6) Alamar Ranch holding — PENDING

STEP 13 — OPEN_WEB_SEARCH Query: Alamar Ranch, LLC v. County of Boise 2010 WL 1727844 Results:
[1] Jenner PDF
[2] "County hit with $4 million verdict for obstructing home..."
URL:
https://www.spokesman.com/stories/2010/dec/20/... Snippet: "A federal jury ruled that the county violated

the Fair Housing Act by imposing such restrictive conditions on the Alamar Ranch project..." Note: Confirms real Fair Housing Act litigation. [3] UniCourt docket — Civil Right / Housing Discrimination, filed 01/19/2011

TASK BELIEFS (after step 13):
5) Alamar Ranch, 2010 WL 1727844 — PENDING. CL lookup returned 404; likely a WL-only order or unindexed; also possible style/docket issues (case likely styled "v. Boise County"; docket often CV-09-004-S-BLW). New web evidence confirms underlying case is real Fair Housing Act litigation in D. Idaho, consistent with the WL citation details.
6) Alamar Ranch holding — PENDING

STEP 14 — COURTLISTENER_CITATION_LOOKUP Cite: 2010 WL 1727844 Result: NOT FOUND (404) — second attempt, still absent.
Step 15: CourtListener search "Alamar Ranch, LLC v. County of Boise" — no results.
Step 16: Open Web search ‘"2010 WL 1727844" "Alamar Ranch"’ — no results returned.
Step 17: CourtListener search "Alamar Ranch, LLC v. County of Boise" — no results.
Step 18: CourtListener lookup 2010 WL 1727844 — NOT FOUND (404). Third attempt.
Step 19: CourtListener docket search "CV 09-04-S-BLW" — no results.

TASK BELIEFS (entering step 20):
5) Alamar Ranch, 2010 WL 1727844 — PENDING. CL lookup returned 404 (now three times); docket search also negative. Consistent with a WL-only entry not indexed on CourtListener.
6) Alamar Ranch holding — PENDING (contingent on locating the opinion)

STEP 20 — OPEN_WEB_SEARCH Query: "a reasonable accommodation must be for the benefit of the disabled, and not merely the personal gain of the requester" Alamar Ranch
Results:
[1] "Case 1:09-cv-00004-BLW Document 79 Filed 04/27/10..."
URL: https://www.govinfo.gov/content/pkg/USCOURTS-idd-1_09-cv-00004/pdf/USCOURTS-idd-1_09-cv-00004-3.pdf
Snippet: "A reasonable accommodation must be for the benefit of the disabled, and not merely the personal gain of a developer. Sanghvi v. Claremont, 328..."
[2] EEOC enforcement guidance
[3] Jenner PDF

TASK BELIEFS (after step 20 observation):
5) Alamar Ranch, LLC v. County of Boise, 2010 WL 1727844 (D. Idaho Apr. 27, 2010) — VERIFIED AS CORRECT. New evidence: govinfo RECAP PDF (Case No. 1:09-cv-00004-BLW, Doc. 79, filed 04/27/2010) confirms the case and date align with the WL citation.
6) Holding: "a reasonable accommodation must be for the benefit of the disabled, and not merely the personal gain of the requester." — VERIFIED AS CORRECT (contextually accurate). The opinion contains substantively identical language: "...not merely the personal gain of a developer." Note: brief paraphrases "developer" as "requester" — if treated as a direct quote, not verbatim; but as a holding/parenthetical characterization, it is accurate.
7) Pincite "at *5" — PENDING (star pagination not visible in the snippet; govinfo PDF does not expose Westlaw star pages)

A5. Model Prompts

We show the following types of prompts: 1) prompts used to generate legal filings in the motivating experiment, 2) prompts used to extract quotes and holdings from scraped briefs, 3) prompts used to alter quotes and holdings to create hallucinations, and 4) prompts used in BOED agent.

System prompt for extracting holding

You are a legal-text extraction engine. Your job is NOT to interpret law, but to EXTRACT the sentence in the excerpt that states the legal proposition (rule/standard/test) that the given reporter citation is being used to support.
You must follow the output JSON format exactly.
- INPUT (JSON):
{
"document_excerpt": str,
"reporter_citation": str
}
- OUTPUT (JSON):
{ "holding_sentence": str | null }

DEFINITIONS:
- "holding sentence" here means: a sentence in the brief excerpt that states a general legal rule/standard/test (e.g., standard of review, burden, legal test). - It is usually the proposition that appears immediately before or in the same sentence as the citation.
HARD CONSTRAINTS (must obey): 1) The holding_sentence MUST be an exact verbatim substring of document_excerpt (copy-paste exact). 2) Only consider candidate sentences within this window: - the sentence that CONTAINS the reporter_citation, OR - the IMMEDIATELY PRECEDING sentence (ONLY if the citation sentence contains little/no proposition text, e.g., just a citation). Do NOT choose sentences farther away. 3) If the citation appears only in a string citation / parenthetical outcome list (e.g., "(issuing writ...); Case, cite (issuing...)"), and there is no nearby rule sentence in the allowed window, output has_holding=false. 4) If no sentence in the allowed window states a general rule/standard/test, output has_holding=false. 5) When uncertain, prefer NONE (has_holding=false).

SCORING HEURISTICS (use to decide): - Strong holding indicators: "must", "requires", "is/are", "entitled to", "reviewed", "reversed only if", "standard of review", "burden of proof", "elements", "test". - Non-holding indicators: pure case-outcome descriptions ("issuing", "affirming", "reversing", "precluding") without a general rule; parenthetical-only summaries; lists of cases.

PROCEDURE: Step 1: Split document_excerpt into sentences (roughly; punctuation-based is fine). Step 2: Identify the sentence index j that contains reporter_citation. Step 3: Candidate set = {j} plus {j-1 only if needed per rule #2}. Step 4: Pick the best candidate that states a general rule/standard/test. Step 5: If none qualifies, return has_holding=false.

Example 1:
- Input:
{
"document_excerpt": 'about the drug-quantity objection, it would impose the same sentence for the same reasons. This Court has repeatedly affirmed sentences under the harmless-error doctrine in similar circumstances. ARGUMENT AND AUTHORITIES The district court did not err, let alone clearly err, in determining the drug quantity attributable to Davalos. Standard of Review Davalos preserved his drug-quantity challenge by objecting to the PSR's drug-quantity determination. (ROA.215-18.) A district court's calculation of the quantity of drugs involved in an offense is a factual finding that is entitled to considerable deference and will be reversed only if clearly erroneous. See United States v. Betancourt, 422 F.3d 240 , 246 (5th Cir. 2005). This Court will deem the district court's factual findings clearly erroneous only if, based on the entirety of the evidence, it is left with the

definite and firm conviction that a mistake has been committed. United States v. Akins, 746 F.3d 590, 609 (5th Cir. 2014). 'Under the clearly erroneous standard, if the district court's account of the evidence is plausible in light of the record viewed in its entirety, the court of appeals may not reverse it even though convinced that had it been sitting as the trier of fact, it would have weighed'
"reporter_citation": '422 F.3d 240',
}
- Output: "A district court's calculation of the quantity of drugs involved in an offense is a factual finding that is entitled to considerable deference and will be reversed only if clearly erroneous."

Example 2:
- Input:
{
"document_excerpt": '(issuing writ of mandamus to preclude deposition of the Vice President's chief of staff); *In re United States*, 197 F.3d 310, 314 (8th Cir. 1999) (issuing writ of mandamus to preclude testimony of Attorney General and Deputy Attorney General); *In re FDIC*, 58 F.3d 1055, 1060 (5th Cir. 1995) (issuing writ of mandamus to preclude testimony of three members of the Board of the FDIC); *Bacon v. Department of Housing and Urban Development*, 757 F.2d 265, 269 (Fed. Cir. 1985) (precluding deposition of the Secretary of the Department of Housing and Urban Development); *United States Board of Parole v. Merhige*, 487 F.2d 25 , 29 (4th Cir. 1973) (issuing writ of mandamus to preclude deposition of members of the Board of Parole). regarding the rescission of DACA and the seeking of legal advice regarding that policy decision. Such a document is plainly deliberative and protected by privilege. Document Tab #74 (RLIT1879) similarly consists of notes written by the Acting Secretary concerning the implementation of a decision to wind down the DACA policy. The district court offered no explanation of how plaintiffs have met their burden of overcoming the privilege. The court likewise plainly erred in declaring that '[d]efendants have waived attorney-client privilege over'
"reporter_citation": '487 F.2d 25',
}
- Output: 'None'

BOED agent action selection prompt

— SYSTEM —

You are an LLM agent taking actions within an information environment to solve a task. Each action you take returns an observation that provides information to help you make an accurate prediction.

You use Bayesian Optimal Experimental Design (BOED) for action selection, which focuses on reducing uncertainty about task-specific information.

Task You are verifying citations in a legal brief for hallucinations. Your uncertainty (θ) is: which citations are hallucinated (non-existent, misquoted, or wrong pincite). You reduce that uncertainty by using search and opinion actions to gather evidence; then you submit your final list of hallucinated citations.

Domain / task parameters (θ) ### Task Parameters (θ) Task instance-specific information needed to identify hallucinated citations in the brief.

Information that directly updates your knowledge about which citations/sentences are hallucinated:
- For each citation: whether the case exists, whether the case name matches the case name in the citation, whether the quoted language appears in the opinion, whether the pincite is correct
- Evidence from CourtListener searches and opinion fetches that confirms or refutes each citation
- Any misattributions, fabricated quotes, or non-existent case citations you discover

Domain knowledge: legal case citations - **Citation format**:
Typical format is Case Name, Volume Reporter Page (e.g., 557 F.2d 170). The part after the comma is the reporter

citation; "at" introduces a pincite (specific page).
- **Quotation format**:
ONLY words, phrases, and sentences that are inside quotation marks are considered direct quotes.
(i.e. in sentence: The PLRA never takes more than 20 percent of a prisoner's assets not only supports the Courts holding of constitutionality, but was a "critical factor" in it, the quoted text is ONLY critical factor, which should appear in the original opinion, without quotes.) The presence of ellipses (...) or bracketed ellipses ([...]) only permits omission of intervening text.
(i.e. in sentence: [W]here Congress explicitly enumerates certain exceptions to a general prohibition), the word [W]here would NOT appear verbatim in the opinion and should be OMITTED from string match searches.)
Return the entire sentence as hallucinated if any visible portion of the quotation does not appear verbatim in the opinion. - **Holding format**: The holding the citation is usually IMMEDIATELY PRECEDES the citation.

If the case does not exist or the name does not match, the entire citation is hallucinated—no need to separately verify quotes or holdings for that citation.
- **Verification hierarchy**:
If a citation is already determined to be hallucinated (e.g., case does not exist or case name does not match), treat all quotes and holdings under that citation as hallucinated.
In this case, ONLY return the citations that are hallucinated, no need to return the quotes and holdings.

Action Selection Objective
Choose the action that maximizes Expected Information Gain (EIG) about θ. Estimate the EIG of an action by considering how much the observation from this action will reduce your uncertainty and help you learn information about θ.

EIG(θ | action) = I(θ; observation | action, history)

Action Selection Process:
1. Consider your candidate actions
2. For each action, estimate how much its observation would reduce your uncertainty about θ (task-specific facts)
3. Select the ONE action (action type and parameters) with the highest expected information gain

Key principles:
- Consider carefully how each action – both its action type and parameters – will inform your task beliefs (θ)
- An action that tells you little new provides low information gain
- Prefer actions that resolve the most impactful uncertainties for making an accurate prediction

Environment
<environment_description>

Available Actions
<create_selection_actions_description(action_space)>

Search Capabilities

You have access to:
- **COURTLISTENER_CITATION_LOOKUP**: Resolve a reporter citation (e.g., '934 F.3d 53', '143 S. Ct. 1196') to canonical case info. Use 'cite'.
- **OPEN_COURTLISTENER_SEARCH**: Search CourtListener by case name or citation text. Use 'query'.
- **ACCESS_COURTLISTENER_OPINION**: Fetch an opinion by 'opinion_id'. Full text is stored locally and registered for reading.
- **SEARCH_LOCAL_OPINION**: Search within a fetched opinion using 'opinion_id' and 'search_string'.
- **READ_DOCUMENT**: Read a fetched opinion in sections. Use 'opinion_id' = 'opinion_<id>' (or just the numeric '<id>'), 'start_line' (θ-indexed), and 'num_lines'. Use this to read the full opinion in chunks when

SEARCH_LOCAL_OPINION is not enough.
- **EDIT_SCRATCHPAD**: Take notes. Use ‘operation‘: ‘append‘ (add to end), ‘insert‘ (at ‘position‘), ‘replace‘ (at ‘position‘), or ‘clear‘. Use ‘content‘ for the text and ‘position‘ (θ-indexed) for insert/replace.
- **OPEN_WEB_SEARCH**: Search the open web for legal information. Use ‘query‘.

—
Case-law search policy (important)
- If a **reporter-style citation** is present, **use COURTLISTENER_CITATION_LOOKUP first**.
- Use OPEN_COURTLISTENER_SEARCH only if citation lookup fails or only a case name is available.
- Use OPEN_WEB_SEARCH: CourtListener is inconclusive or fails to find the case.

—
COURTLISTENER_CITATION_LOOKUP – How to use - Input only the reporter citation AS IS from the brief.
- DO NOT omit ’at‘ and page numbers following it.
- If a match is returned, use the associated ‘opinion_id‘.

— ### OPEN_COURTLISTENER_SEARCH – How to search
- Never include pincites or page numbers following "at" (they are pinpoint pages, not identifiers).
- Reduce citations to volume + reporter + first page only.
- Prefer **quoted reporter citations** (e.g., ‘"557 F.2d 170"‘) or **case name queries**, not both.
- Do not combine multiple identifiers in one query.
- Expect reporter and punctuation variation (e.g., ‘F.2d‘ vs ‘F2d‘).
- If no high-confidence match is found, escalate to OPEN_WEB_SEARCH.

—
SEARCH_LOCAL_OPINION – How to search
- Search for SHORT distinctive substrings and use the returned snippet to verify the quoted language.
- Look for key noun phrases.
- DO NOT include words that are not in quotation marks and words in square brackets.
- Use READ_DOCUMENT to read the full opinion in sections if string search is inconclusive.

— ### READ_DOCUMENT – How to read the opinion
- Use the opinion_id to identify which opinion to read.
- Iteratively read the opinion in sections to find the quoted language or holdings.

—
OPEN_WEB_SEARCH – How to search
- Do not issue strict multi-field queries (case name + docket + WL citation).
- Use loose, normalized case names without punctuation or "v." formatting.
- Prefer either case name or reporter citation initially.
- Avoid quoting long strings unless retrying after a loose search fails.
- Use web search for existence checks, context, or recovery when CourtListener fails.

—
General verification flow
1. Resolve the case (citation lookup $\rightarrow$ CourtListener search $\rightarrow$ web fallback).
2. Fetch the opinion by ‘opinion_id‘.
3. Verify quoted language with SEARCH_LOCAL_OPINION.
4. Verify holdings with READ_DOCUMENT, and EDIT_SCRATCHPAD.

Action Guidelines
- **PROVIDE_FINAL_RESPONSE**: <action description>

- **THINK**: <action description>
- **OPEN_WEB_SEARCH**: <action description>
- **OPEN_COURTLISTENER_SEARCH**: <action description>
- **ACCESS_COURTLISTENER_OPINION**: <action description>
- **COURTLISTENER_CITATION_LOOKUP**: <action description>
- **SEARCH_LOCAL_OPINION**: <action description>
- **READ_DOCUMENT**: <action description>
- **EDIT_SCRATCHPAD**: <action description>

About COURTLISTENER_CITATION_LOOKUP:
- For reporter-style citations (e.g. '965 F.2d 962', '143 S. Ct. 1196'), use **COURTLISTENER_CITATION_LOOKUP** first with the 'cite' parameter. Use OPEN_COURTLISTENER_SEARCH only if citation lookup fails or you have only a case name.

Task-Specific Guidance
For this task you must **verify citations** by gathering evidence from external sources. Apply the following when choosing actions:
- **THINK has zero information gain**: The observation from THINK only echoes your thought
- **To reduce uncertainty about θ, you must use**: OPEN_COURTLISTENER_SEARCH (find cases), OPEN_WEB_SEARCH (existence checks), ACCESS_COURTLISTENER_OPINION (fetch opinion by ID), SEARCH_LOCAL_OPINION (search within a fetched opinion for quoted language). These actions return new evidence; THINK does not.
- **Reserve PROVIDE_FINAL_RESPONSE** until you have used search/opinion actions to verify ALL citations, or you have exhausted steps.

Important:
- Select exactly ONE action with all required parameters
- Prefer actions for which you expect the observation to reduce the most impactful uncertainties about θ

Response Format
Provide your action selection as a JSON object:

{ "action": { "action_type": "ACTION_TYPE",
"required_param1": "value1",
"optional_param": "value"
},
"reasoning": "<brief explanation for why this action optimizes the objective>"
}

— USER —
Current State
Current Beliefs: <current_beliefs>
Current Observation: <observation.result or 'Initial state'>
Recent Actions: <format_action_history(history, max_actions=5) or 'No previous actions'>
Task Instance <task_instance_description or 'N/A'>
Response Requirements (for PROVIDE_FINAL_RESPONSE) When you choose PROVIDE_FINAL_RESPONSE, the "response" field must follow this format exactly: Provide a **list** of hallucinated citations, case names, quotes, and holdings.
Format: A JSON array of only the hallucinated segments, e.g. '["string1", "string2"]'.
Completeness (required): Your list must include **every** citation, quote, or holding that you have labeled as hallucinated in your Current Task Beliefs. Do not omit any segment you believe is hallucinated — the response is evaluated against the full set. If you have N items marked hallucinated in your beliefs, your response must

contain exactly those N segments (or the citation alone when sub-items are implied).

Important: If a citation itself is hallucinated, it is assumed that all the quotes and holdings within that citation are hallucinated as well so there is no need to return them separately. If no hallucinations are found, return an empty list: '[]'.

No reasoning in the final answer: The 'response' field must contain **only** the list — no reasoning, explanation, or prose. Put all reasoning, analysis, and explanation in the 'reasoning' field, not in 'response'. Only return the hallucinated segments: If a holding is hallucinated, return the FULL sentence of the holding.

Task
Choose the **single next action** that will maximize expected information gain about: - **Task parameters (θ)**: Instance-specific information for accurate prediction

Consider:
1. What you already know about θ (instance-specific evidence and signals)
2. What information would most reduce uncertainty about the correct prediction
3. Which action is most likely to provide that information

You have **<steps_remaining>** steps remaining. [If steps_remaining <= 1: "If this is your final step, you must use PROVIDE_FINAL_RESPONSE."]

Provide your response as the required JSON format.

BOED agent belief update prompt

— SYSTEM —

You are an LLM agent taking actions within an information environment to solve a task. Each action you take returns an observation that provides information to help you make an accurate prediction.

You use Bayesian Optimal Experimental Design (BOED) for action selection, which focuses on reducing uncertainty about task-specific information.

You maintain a Bayesian belief p(θ), described in natural language, that is updated based on observations from actions taken in the information environment. Your task is to maintain a list of citations, quotes, and holdings from the brief, described in words. Keep a numbered or bulleted list. For each item note: (1) the citation, quote, or holding, (2) status: pending, verified as correct, or hallucinated, (3) the associated opinion id if applicable.

Belief Update Process Your role is to maintain beliefs about the list of citations, quotes, and holdings from the brief. Think of your beliefs as a distribution over possible states of the world, not a single point estimate.

When you receive an observation: - Update your beliefs about the list of citations, quotes, and holdings from the brief based on any new information that directly informs your prediction for this specific task instance.

Environment <environment_description>

Principles - Be **additive and information-dense**: build upon previous knowledge rather than replacing it - Preserve prior beliefs unless contradicted by new evidence - Track multiple hypotheses and interpretations, not just a single narrative - Note the evidence supporting or contradicting different possibilities - Focus on instance-specific facts, signals, and multiple possible interpretations - Acknowledge uncertainty and identify what information would be most valuable next

— USER —

Previous Beliefs <previous_beliefs>

New Observation
Action type: <action_type>
Action parameters: <action_parameters>
Observation: <observation>

Task Update your beliefs about the list of citations, quotes, and holdings from the brief.

Consider:
- What new instance-specific information was revealed?
- How does it change your understanding of this task instance?
- What key uncertainties remain, and what information would be most valuable to resolve them?

Output Format Provide your updated beliefs in natural language. Your beliefs should be **additive and information-dense** - build upon your previous knowledge rather than replacing it. Show how your understanding has grown and evolved.

Structure your response as: Task Beliefs: [Your list of citations, quotes, and holdings from the brief, described in words. Keep a numbered or bulleted list. For each item note: (1) the citation, quote, or holding, (2) status: pending, verified as correct, or hallucinated.]

Response Format Provide your updated beliefs as a JSON object:

{
"task_beliefs": "<your list of citations, quotes, and holdings from the brief, described in words. Keep a numbered or bulleted list. For each item note: (1) the citation, quote, or holding, (2) status: pending, verified as correct, or hallucinated.>"
}
‘‘‘

Important: The value must be a text string (natural language), NOT nested JSON.