

"Hallucination" Is Not Enough: Why Legal LLM Benchmarks Should Measure "Correctly Grounded"

Adam Piotrowski ([ORCID 0009-0005-7144-6047](https://orcid.org/0009-0005-7144-6047)) · LexAlpha Lab, KTZR · 26 October 2026 · working paper, Version 1.2 · DOI: [10.5281/zenodo.23438195](https://doi.org/10.5281/zenodo.23438195) (all versions)

Changes from v1.1 (17 October 2026): the text of §1–§8 and §11–§12 is edited for plain English, without changes to content or numbers, except for one added paragraph in §2 on the distinction between hallucination and factuality [17] and one in §6.2 on the extended W0/W1 set; two sections are added: §9 on the accuracy of model judges and on the verification report, and §10 on OKF-Legal; the former §9 and §10 become §11 and §12; Tables 1–9 and Figures 4–9 are added to the text, and Appendices A–G (systems and data sets, scoring specification, question set, model judges with confusion matrices and intervals, verification layer, error classes, cost) replace the former Appendix A; references [11]–[17], an ORCID and a disclosure on OKF-Legal are added; the statements on data availability and the use of AI are updated. All figures are recomputed from the reference cards as of 25 October 2026.

Changes from v1.0 (16 October 2026): the author's final review decisions on three scoring cards are applied; the in-house system of the author's firm, run in two configurations, is excluded from the comparison (13 systems, 65 answers instead of 15 and 75); all numbers are recomputed from the cards; error types in Figure 3 are counted per answer; the list of systems is added as Appendix A; statements on data, competing interests and the use of AI are added.

Abstract

Legal benchmarks of large language models (LLMs) commonly report a "hallucination rate". A widely cited operationalisation, by Magesh et al. (2025), decomposes factual hallucination into *correctness* and *groundedness* and labels a response as hallucinated when it is incorrect or misgrounded. This paper argues, on the basis of a working benchmark of LLMs on Polish civil procedure (LexAlpha Lab), that the framework does not apply to civil-law systems without modification. Three observations support the claim: (1) the two dimensions overlap on the error types that matter most in statute-based law; (2) a wrong case citation and incorrect content attributed to a statutory provision do not carry the same weight where the statute, not the judgment, is the source of law; (3) the currency of the provision's wording, which Polish official sources (ISAP) date precisely, is not measured by any of the empirical benchmarks reviewed here. I propose to replace the binary question "did the model hallucinate?" with "is the answer **correctly grounded**?", defined as: the cited designation exists, the source supports the content attributed to it, and the wording was in force on the legal-state date of the question. Preliminary results from 65 answers (10 models without tools, 3 systems with search; 4 model judges, Cohen's κ 0.81–0.93 against the cards before the author's review) are reported, together with the limitations of the pilot and the reasons for keeping the criteria simple enough to re-run on every monthly model release. Version 1.2 adds a test of the model judges on cards closed before any judge ran: the judges, all with web search, are accurate on errors that can be detected with the answer, the answer key and the scoring rules, and inaccurate on errors that require a source which web search does not reliably return (stale wording, a real case number attached to a thesis from another case). A verification report produced from the official sources by citation-checking code and a model-based support check, given to the judge as auxiliary evidence raised the

agreement of one judge with the lawyer's cards (Sol 6.1, κ 0.84 $\rightarrow$ 0.97) and lowered that of another (Gemini 4 Argon, κ 0.95 $\rightarrow$ 0.88); the same report used as binding evidence lowered the agreement of both. A third variant, which shows a verification status instead of "✓" whenever the wording on the legal-state date was not established, left the agreement of one judge unchanged and lowered that of the other relative to the second variant.

Keywords: legal hallucination · groundedness · LLM evaluation · civil law · Polish law · LLM-as-a-judge · ISAP

1. Introduction

When we started work on a legal benchmark, I took the Stanford paper "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools" [1] as the reference framework. Its definition of hallucination seemed precise, and I assumed that it was universal.

After several months of building our own benchmark for LLMs in Polish law, I no longer think so. The definition is not wrong. But in civil law, and in Polish law in particular, it does not cover every case, and in practice it merges two different things: whether the model recalls the right legal basis for a statement, and whether the statement itself is supported by that basis.

As far as I could establish, there is no broad public benchmark of legal hallucination in Polish law, covering several branches of law and many systems, with a published sampling and annotation protocol. There are two rulings of the Supreme Administrative Court (NSA) from 2026 that name "AI hallucinations" in lawyers' briefs [8], and one study of the public procurement (KIO) exam in which a GPT-4o judge scored the answers 51 to 77 points higher than the human examination board [9]. The second finding shows why the scoring method matters.

2. The Stanford framework

Magesh et al. [1], following Dahl et al. [2], list three ways in which a model can hallucinate: it can be unfaithful to its training data, to the prompt, or to the facts of the world. Their paper deals with the third. Other work draws the line elsewhere. Bang et al. [17] separate "hallucination", defined against the model's training data and the input, from "factuality", defined against an external source of truth that changes over time. Under that definition a model that gives an earlier wording of a statute, one it saw in training, does not hallucinate; it makes a factuality error. In legal answers this is a frequent error (class W, §6), and a benchmark that reports only a hallucination rate in this narrower sense would not count it.

In the framework of Magesh et al., factual hallucination has two dimensions: *correctness* (the response is factually correct and relevant) and *groundedness* (the relation between the response and the sources it cites). A response is hallucinated if it is incorrect or misgrounded. One detail matters for the argument below: in that framework a correct answer with no citation is not a hallucination, only "incomplete". The authors also stress that "grounded" in their legal sense means grounded in the governing law, not in the document that the system retrieved, which is what the word means in computer science.

3. Three ways models ground their answers

Groundedness is the dimension that needs closer examination. In our evaluations we see three grounding practices:

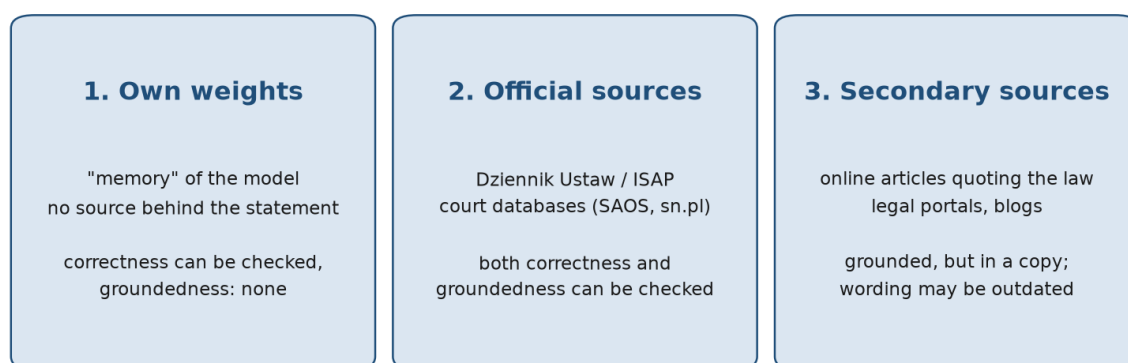
1. grounding in the model's own weights, its "memory";
2. grounding in official state sources of law (in Poland: Dziennik Ustaw, ISAP, court databases such as SAOS);

3. grounding in unofficial sources, such as online articles that quote the law.

The third practice is "grounded" only in the computer-science sense. Our scoring cards record it with a separate label, "**no source verified**", with three values: *no tools*, *secondary sources only*, *official source reached*.

How LLMs ground legal answers - three practices

observed in the LexAlpha Lab working benchmark, Polish civil procedure



Working rule: an answer is correctly grounded only if designation, content and wording all match the official text.

LexAlpha Lab · working paper v1.0 (pilot)

Figure 1. Three grounding practices observed in the LexAlpha Lab benchmark.

4. Edge cases: one category in the paper, four distinct problems for a lawyer

Some cases fall into one category in the Stanford framework ("misgrounded") but are different problems for a practitioner:

- the model recalls the right content of a statute or judgment but gives it a wrong designation (the correct case number is I ACa 234/22, the model cites I ACa 123/22);
- the model gives the right designation but wrong content;
- designation and content are right in substance and wrong in one detail, such as art. 1 instead of art. 1¹ (common in Polish law, where amendments add dozens of units inside one article);
- the model quotes an old wording of a provision as the current one;
- for some products we cannot establish whether web search or a database was used, so groundedness cannot be assessed.

The first four are error types; the fifth is not an error in the answer but a gap in what we can observe about the model's process.

Correctly grounded? One bucket in the paper, four problems for a lawyer

designation of the source (article, case number) × content attributed to it · only the top-left cell is correctly grounded

	content right	content wrong
designation right	designation right content right the only clean case (still: official or secondary source?)	designation right content wrong statement not in the provision most frequent error in our cards
designation wrong	designation wrong content right I ACa 123/22 Instead of I ACa 234/22 art. 1 Instead of art. 1 ¹	designation wrong content wrong fictitious source the classic hallucination

Plus a fifth case: outdated wording of a real provision presented as current law.

And a sixth: no way to tell whether a tool searched the web or used a database - there groundedness cannot be assessed at all.

Figure 2. Designation × content. Only the top-left cell is correctly grounded; in the Stanford framework the other three cells would fall under "misgrounded", although that procedure assesses groundedness only for responses already judged correct.

In these cases groundedness overlaps with correctness. The paper's own example, "the right to same-sex marriage is protected, *Miranda v. Arizona*", is coded as correct but misgrounded. Take the opposite case: a correct article number with content that is not in the provision. The response is both incorrect, because it makes a false statement about the law, and misgrounded, because the cited source does not support the claim. The paper settles this by the order of assessment: groundedness is assessed only for correct responses, so correctness is decided first. For the final label the order does not matter, because an incorrect or a misgrounded response is a hallucination either way. I conclude that the two dimensions should be scored and reported separately. In our scoring cards the substance of the answer and the grounding of each citation are two separate scores.

5. Common law and civil law: a wrong citation does not have the same weight

There is a second reason why the framework does not apply to Poland without modification. It was built for US law, where case law is a source of law and a wrongly cited judgment is the central error; "inapplicable authority" (wrong jurisdiction, wrong court, overruled case) is one of the main causes of hallucination in the paper's typology. In a civil-law system the hierarchy is different. The content of the provision comes first; a judgment is usually an interpretation of it. In our scoring cards both errors give A6 = 0, but a wrong case number with a correct thesis is recorded as a peripheral error, and a correct article number with content that is not in the provision as an error in the core (Appendix B.6). Both are "misgrounded" in the Stanford terminology. The weighting is a judgement about harm: for a Polish lawyer the second error is, in my assessment, the more dangerous one. It is also the most frequent class in our results (class T, which covers provisions and judgments alike).

Two studies from civil-law countries looked at a narrower question: do models cite judgments less reliably than statutes? On the Japanese bar exam, an annotator with legal training marked 295 of 441 precedent citations as hallucinated (67%), but only 381 of 2,575 statute citations (15%). The authors conclude that the models' knowledge of precedents is weaker than their knowledge of statutes. One caveat: in one of their prompt settings the models were given the relevant statutes, which lowers the statute error rate [5]. On a Brazilian legal platform, 53% of case-law references could not be matched to an official database, against 18% of references to legislation. The authors explain this mainly by the difficulty of normalising case-law citations. The figure

therefore says whether a reference can be found, not whether the model's claim was correct [6]. Neither study compares the harm of the two kinds of error. Our scoring records a wrong case number with a correct thesis as a peripheral error. That is a judgement about what harms the lawyer, and these two studies neither support nor contradict it.

Table 1. Two civil-law studies on citations of judgments and statutes.

Study	Setting	What was measured	Judgments	Statutes
Choi et al. 2026 [5]	Japanese bar exam, written answers, one annotator with legal training	citations marked as hallucinated (source inconsistent with the actual law or precedent)	295 of 441 (67%)	381 of 2,575 (15%)
Figueiredo et al. 2026 [6]	Brazilian legal platform, 43,175 references in production	references that could not be matched to an official database	53%	18%

6. Method and preliminary results

6.1 Setup

The pilot run (r3) uses 5 questions from Polish civil procedure (service through the court portal, rejection of an interlocutory appeal, "principal reasons" versus written justification, enforcement clause against a spouse, order of payment on a bill of exchange versus security). Each question has an answer key: the core of the answer, required legal bases, rubric elements, penalty elements and the legal-state date.

All answers were collected on 10 October 2026; the legal-state date in every answer key is 1 October 2026. Questions were asked in Polish, with the same short system instruction ("You are an assistant to a lawyer in a Polish law firm...") wherever the interface allowed one. Ten models from three vendors (OpenAI, Anthropic, Google) answered without tools, without search and without access to the texts of statutes; 3 systems answered with web search. The full list with model identifiers and settings is in Appendix A. The pilot also included the in-house research assistant of the author's firm, in two configurations; both are excluded from all numbers in this paper because of the conflict of interest. For each answer a scoring card is filled in with: core (A-R), required legal bases (A2), quotations (A3), hallucination (A6, binary), framing (A8), rubric (R), penalties (K) and a usefulness block (B). The full specification, the cards and the question set are described in the Polish working report [7] and summarised in Appendix B.

A6 is binary. An answer receives 0 and the label *hallucinated* when at least one statement about a source of law is inconsistent with the official text on the legal-state date. The class of error is recorded separately: **F** (the source or unit does not exist), **T** (the cited provision or judgment is attributed content which is not there), **W** (wording from another date presented as current), **S** (a real source with a distorted identifier). An answer without any citation is not a hallucination; it loses the points for legal bases and is flagged "no source verified".

Every card is first filled in by a Claude model, with every cited provision checked against the official text in ISAP and SAOS, and then re-scored blind by four model judges: Sol 6.1 and Astra 6 (OpenAI, via Codex) and Gemini 3.1 Pro and Gemini 3.8 Flash (Google, Vertex AI, with Google Search). Three of the judges (Sol 6.1, Astra 6, Gemini 3.1 Pro) are also among the evaluated models, so in some cards a judge scores its own answer. A human (the author) reviews the cards on which the judges disagree with the card.

6.2 Results

Under the binary rule **41 of the 50 answers given without tools were labelled hallucinated**; for 7 of the 10 models all five answers carried the label. The three systems with web search had the label in 2 of 15 answers (ChatGPT in a temporary chat 0 of 5, ChatGPT in a new chat 1 of 5, Sol 6.1 with search 1 of 5). Among the 41 hallucinated answers given without tools **the most common error was not a fictitious source but a correct designation with content which is not in the provision** (class T), found in 35 answers. Old wording (class W)

was second with 13 answers; a fictitious source (class F) appeared in 4 answers and a real source with a wrong identifier (class S) in 1.

On grounding, of the 15 answers from systems with search, 12 reached an official source and 3 used only secondary sources (legal portals, blogs). With the 50 answers given without tools included, in 53 of 65 answers in the whole run the model worked only on its own weights or on secondary sources.

Table 2. Pilot run: hallucination and source access by condition (65 answers, 5 questions).

Systems	Answers	Hallucinated	Models that failed on all 5 questions	Official source reached	Secondary sources only	No tools
10 models without tools	50	41	7 of 10	0	0	50
3 systems with web search	15	2	0 of 3	12	3	0
All	65	43	—	12	3	50

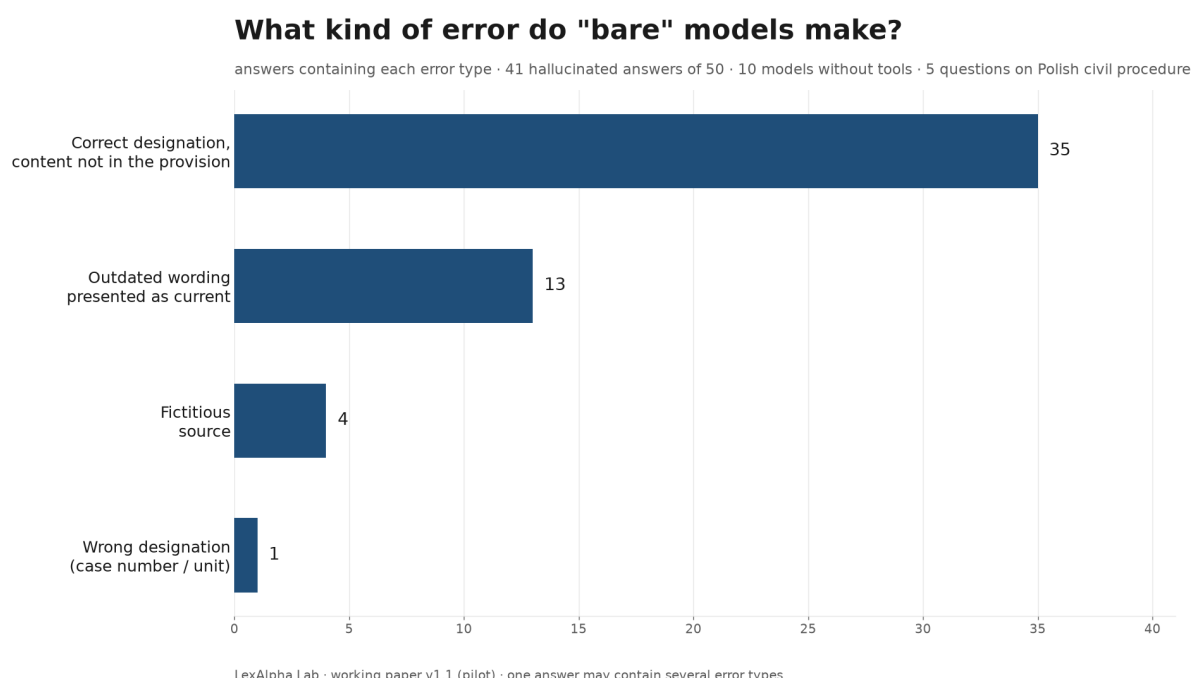


Figure 3. Error classes in 41 hallucinated answers from 10 models without tools: number of answers containing each class (one answer may contain several).

After the pilot the benchmark was extended to ten questions and every model answered each question twice, without tools and with web search (Appendix A.2). The 160 answers are: 30 answers without tools to the pilot questions from the six pilot models that continued, re-scored under the current cards; 10 new answers without tools from Gemini 4 Argon and Gemini 3.8 Flash; 40 answers with search to the pilot questions; and 80 answers to the five new questions. Under the current specification, of those 160 answers 91 carry the label "hallucinated" (51 of 80 without tools, 40 of 80 with search) and 46 are correctly grounded (19 and 27). Classes T and W remained the two most frequent: class T in 66 answers (42 without tools, 24 with search), class W in 35 (18 and 17), class S in 9 (4 and 5), class F in 8 (4 and 4); one answer may contain several classes. The comparison is paired: the same eight models answered the same questions in both conditions. In this set the search condition had fewer class T answers and about the same number of class W answers. What differs between conditions is not only the presence of search but also how each provider implements it (Appendix A. 2), so the effect is that of each provider's search setup, not of search in general. The remaining 23 answers (10

without tools, 13 with search) contain no false statement about a source but do not satisfy all core-answer criteria: 20 satisfy some of them and 3 none, one of the 3 because the provider blocked the answer. The figures are given by condition and by class, not by model, because the set is a calibration set; details are in Appendix F.

Table 3. The W0/W1 set: labels and error classes by condition (160 answers, 10 questions, 8 models; cards of 25 October 2026).

	W0: no tools (80)	W1: web search (80)	Total (160)
hallucinated (A6 = 0)	51	40	91
correctly grounded (A6 = 1, full core, at least one citation checked)	19	27	46
no false statement, core incomplete or missing	10	13	23
class T: content not in the source	42	24	66
class W: wording from another date	18	17	35
class S: distorted identifier	4	5	9
class F: source or unit does not exist	4	4	8

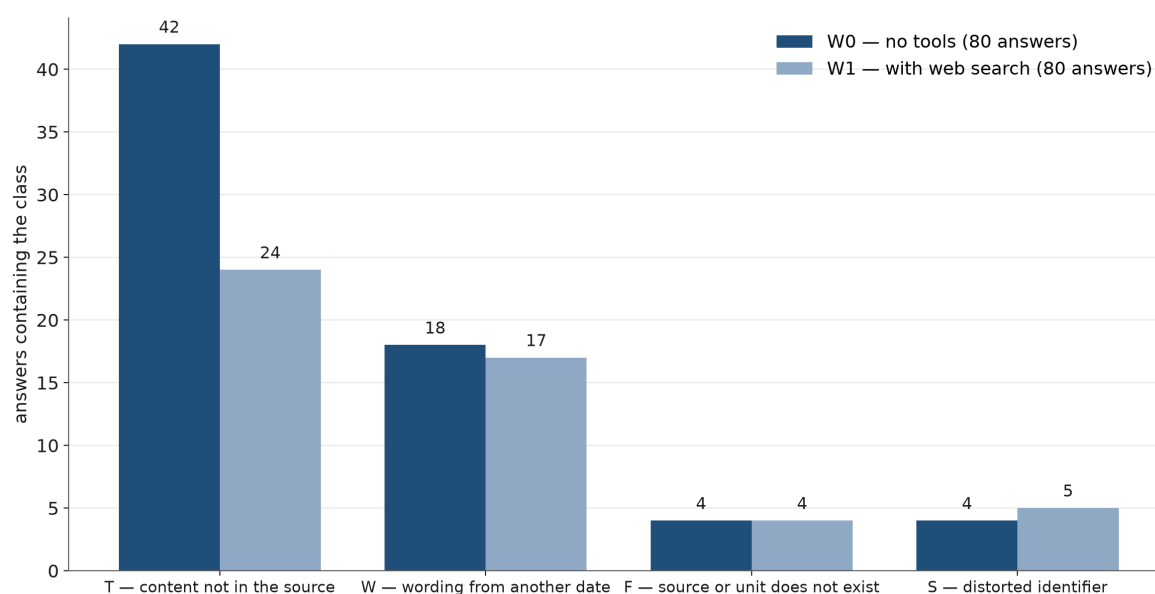


Figure 4. Answers containing each error class, without tools (W0) and with web search (W1), in the W0/W1 set. One answer may contain several classes.

On the hallucination label the model judges agree with the cards at Cohen's κ 0.81 to 0.93 per judge (Astra 6 0.93, Sol 6.1 0.92, Gemini 3.1 Pro 0.84, Gemini 3.8 Flash 0.81), measured before the author's review. Magesh et al. report κ 0.77 between their human labellers. The two figures are not directly comparable: the first scorer of every card is itself a model, so our figure is agreement between models, checked by a human only where they disagree.

6.3 Limitations

The pilot run uses five questions from one branch of law, 13 systems, a single run per system, and model judges validated against the cards rather than against an independent human sample. Three judges are also evaluated models. The author's review changed three cards, all of them cards on which the judges disagreed with the card. κ is therefore reported before the review; after the review the agreement is partly the result of the author accepting the judges' view (on the current cards it is 0.83–1.00, §9.1). One system (ChatGPT in a new

chat) was scored by one judge only, and the model behind the ChatGPT app is not disclosed by the interface. A validation sample scored by human lawyers is planned.

7. From "hallucination" to "correctly grounded"

I therefore propose to replace the question "did the model hallucinate?" with **"is the answer correctly grounded?"**. Magesh et al. use this phrase: grounded "in the legal sense", in the governing law. In our scoring cards every cited provision is checked against the official text. The answer is correctly grounded only if **the designation is correct, the content is present in the source and the wording is current**. A single wrong unit (art. 1 instead of 1¹) fails this test, even if the substance of the answer was correct. Among the benchmarks reviewed in this paper, none checks that the wording was in force on the legal-state date of the question; ISAP provides that date at no cost.

A correct answer without any source is not a hallucination, but it is not correctly grounded either. In the cards an answer counts as correctly grounded when A6 = 1, the core is complete and at least one citation was checked against the official text; the 46 answers in Table 3 all meet the third condition.

This is a strict standard. The standards applied to human lawyers do not transfer to models: a lawyer who writes § 1¹ instead of § 1² has made a typing error and knows what was intended; a model has no such record of intent.

The Stanford paper gives one more reason to keep the "no source" label separate. The authors observed Westlaw stating propositions from overruled cases without citing them, probably because "red-flagged" citations were suppressed in a second pass. A statement without a source cannot be checked against a source, only against the law as independently established, and in that example it was exactly where the suppressed, overruled authority ended up. That is the reason to keep "no source verified" as a separate status rather than fold it into either "correct" or "hallucinated".

8. Why the criteria have to be simple

The Stanford study was conducted by a university team with a preregistered dataset of 202 queries, human raters and a protocol which took months: the dataset was registered in March 2024, the runs took place between March and May, the preprint came out at the end of May and the journal version in 2025. The result is reliable, but it describes Lexis and Westlaw as they were in spring 2024.

New models are released almost every month. If an evaluation takes months, its results describe models which are no longer in use. To keep a benchmark current, the criteria must be clear, fast to apply and repeatable on every new model, even at the cost of some detail. **This is why our rule is binary**: not because it is the most precise rule, but because it can be applied to a newly released model with a documented procedure. Under the binary verdict the cards keep per-citation labels (which provision, which error class), so finer rates can be computed without changing the method.

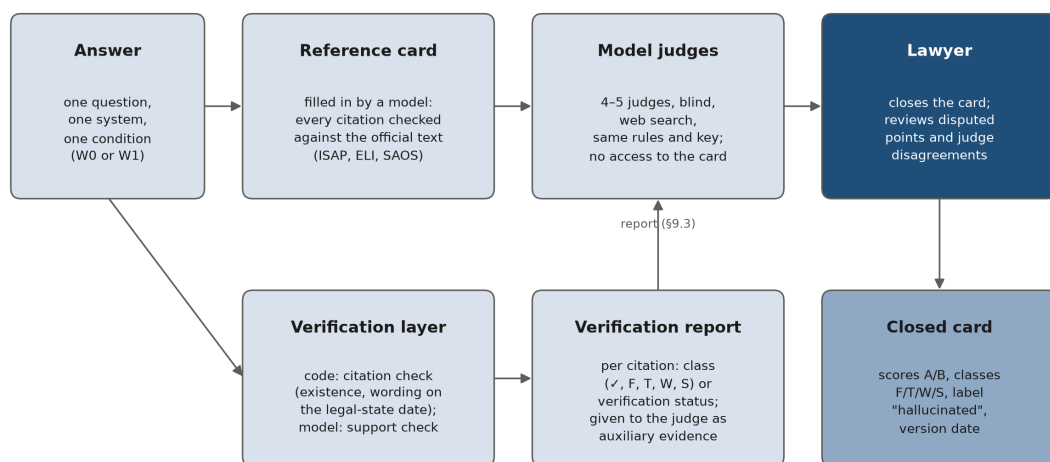
Table 4. Two evaluation protocols.

	Stanford study [1]	This benchmark (W0/W1 set)
Material	202 queries, 4 products	10 questions × 8 models × 2 conditions = 160 answers
Scoring	human raters, per response	reference card per answer, filled in by a model and closed by one lawyer; per-citation classes F/T/W/S
Judges	—	4–5 model judges with web search, blind to the card; κ reported against the cards
Preparation of the reference material	registered dataset, March 2024	answer keys and reference cards: 2–3 weeks of the lawyer's time for 160 answers

	Stanford study [1]	This benchmark (W0/W1 set)
Scoring pass	human raters; runs March–May 2024, preprint May 2024, journal 2025	model judges: hours; verification layer: about 15 minutes of code plus a model pass
Model cost of one scoring pass	not reported	about USD 24–32 for 160 judge cards with the verification report (tokens and searches), excluding the production of the reports and all human work (\$9.4)
Result describes	Lexis and Westlaw as of spring 2024	the models listed in Appendix A as of October 2026; repeatable on a new release

Other groups take the same direction. LegalHalBench [3] counts a cited statute as non-hallucinated only if its name, number and content all match the reference, and a 2026 position paper by Taranukhin and Shwartz [4] proposes to measure legal hallucination as a failure of "legal warrant": the authority must exist, fit the jurisdiction, be current for the date of analysis and actually support the claim (see §9 and §10).

On one point we depart from Magesh et al. They write that letting AI tools "check themselves" is not suitable for legal evaluation and that responses must be hand-scored by lawyers. We use model judges, as described in §6.1, and we report their agreement with the cards. I do not claim that this replaces human scoring. It is the cost of keeping up with monthly releases. Our aim is a system in which a human reviews only a small share of the cards: those on which the judges disagree or the risk of error is highest. Without that, the time of the human reviewer sets the pace of the whole benchmark. This concerns the scoring of a benchmark, not legal practice. In practice every citation still has to be checked, as the SRA and CCBE guidance require [10].



Specification v1.0.1 with recorded decisions (Appendix B). The lawyer sees every card; model judges and the verification layer are measured against the closed cards.

Figure 5. The path of one answer through the evaluation. The lawyer sees every card; the model judges and the verification layer are measured against the closed cards.

9. Can an LLM be a reliable judge of other LLMs' answers?

In §8 I argued that model judges are the cost of keeping a benchmark current. The remaining question is whether they are accurate. The disagreements between the judges and the reference cards fall into two groups. On the part of the evaluation that can be done with the answer, the answer key and the scoring rules, the judges rarely disagreed with the cards. **Most of their misses concerned the part that requires the correct source:** the official text of a provision as in force on a given date, the full text of a judgment, the line of case law. All judges in this study had web search and were instructed to verify every citation in the official sources. Most of

the errors they missed are errors that web search does not reliably expose. This is not a problem of benchmarks alone. The same failure occurs at the retrieval stage of legal assistants: the model receives an incomplete or outdated text and reasons correctly from a wrong premise. We observed both failures in this study. Some judges searched little and missed errors that a search would have revealed, and several official sources we used were truncated or could not be dated. "Garbage in, garbage out" is usually quoted as a slogan. Our results give it a form that can be measured: without better retrieval and better quality of the sources provided, improvements in the models' legal reasoning are hard to demonstrate, because many of the errors that remain are errors of the source rather than of the reasoning. Our design does not separate the two causes; the observation is that source retrieval and version identification contributed to most of the errors the judges missed. An end-to-end benchmark of legal retrieval-augmented generation reached a similar conclusion: retrieval sets the upper limit of answer quality [16].

9.1 Model judges compared with the reference cards

Procedure. Every scoring card in the benchmark was first filled in by a Claude model, with each citation checked against the official source, and then reviewed by the author. The author decided every disputed point and closed the card (§6.1). Disputed points were of two kinds: borderline cases flagged by the model that filled in the card, and disagreements between the card and the model judges. In the pilot set the author saw the judges' scores before closing the cards; in the extension set the cards were closed before any judge ran. Rules were still refined after the judges ran; where a later rule change altered a card's label (one card, 23 October), the current card is used and the case is noted below. On the cards as first closed the extension-set figures were κ 0.92, 0.82, 0.78 and 0.56 for the four judges; on the current cards they are 0.90, 0.80, 0.80 and 0.58. The closed cards are the reference in this section; the procedure for closing a card is described in [7]. Each model judge then scored the same answers independently.¹ For each answer the judge received three inputs in one prompt: the scoring rules (blocks A and B of the specification, including the boundary examples and the decisions recorded in it), the answer key of the question in full (core answer, required provisions, reference wording of the provisions, rubric, penalties, expected calibration, legal-state date), and the text of the answer without the name of the system that produced it. The judge did not see the reference card or the other judges' outputs. The instruction required the judge to check every cited provision and judgment in the official sources by web search (ISAP and ELI for Polish acts, EUR-Lex for EU law, SAOS or the Supreme Court website for judgments), to compare the unit number and its content literally, not to rely on memory, and to list in a separate field every citation it could not check. The judge returned a structured card with the same fields as the reference card: scores for each criterion, the class of each error (F, T, W, S), the binary label "hallucinated", a list of checked citations with their source and verdict, and a short justification. We measured Cohen's κ between each judge and the reference cards on the label "hallucinated".

Agreement. On the 65 cards of the pilot run the four judges of v1.1 were close to the reference on the current cards (κ 0.83–1.00: Astra 6 and Sol 6.1 1.00, Gemini 3.1 Pro 0.93, Gemini 3.8 Flash 0.83). These figures are therefore inflated. A stricter test is the extension set: five new questions, 80 answers, cards closed before any judge ran. There the results diverged: **Gemini 4 Argon κ 0.90** (79 cards; one card without a verdict), **Sol 6.1 0.80, Gemini 3.1 Pro 0.80, Gemini 3.8 Flash 0.58**. With five questions the 95 % intervals are wide (Appendix D.5). The intervals of Gemini 4 Argon (0.80–0.97) and Gemini 3.8 Flash (0.40–0.70) do not overlap; the others do. Separate intervals do not test a difference directly, because the judges scored the same answers; a paired comparison would need more questions.

Table 5. Model judges on the extension set (80 answers, 5 questions, cards closed before any judge ran).

Judge	n	κ	95 % interval	False alarms	Misses	Searches per card	Cards without search	"Unchecked" citations per card
Gemini 4 Argon	79	0.90	0.80–0.97	1	3	12.2	0	0.01
Sol 6.1	80	0.80	0.59–0.95	7	1	11.2	0	0.36

Judge	n	κ	95 % interval	False alarms	Misses	Searches per card	Cards without search	"Unchecked" citations per card
Gemini 3.1 Pro	80	0.80	0.75–0.85	0	8	4.3	25	0.00
Gemini 3.8 Flash	80	0.58	0.40–0.70	1	16	4.7	9	0.01

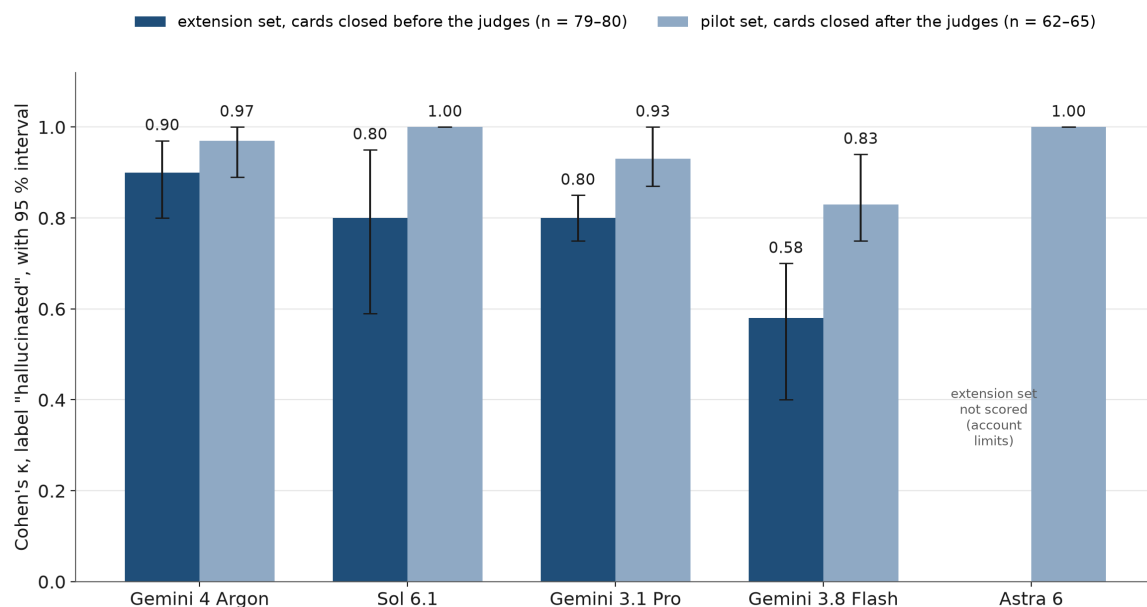


Figure 6. Agreement of the model judges with the reference cards on the label "hallucinated", with 95 % intervals from a bootstrap over questions (Appendix D.5). On the pilot set the cards were closed after the judges' outputs were seen. For comparison, the automatic claim verifier in HalluLens agreed with human annotators on 76.8 % of claims [17]. That figure is raw agreement on single claims, not κ on whole answers, so the two are not directly comparable. The errors were not symmetric. The Gemini judges rarely reported a hallucination that the reference cards did not record; they failed to report hallucinations recorded in the reference cards (Gemini 3.1 Pro missed 8 and Gemini 3.8 Flash 16 of the 42). Sol 6.1 did the opposite: it reported seven hallucinations in answers that the cards had accepted, and missed one. In six of the seven Sol 6.1 applied a stricter definition than ours; in none did it misstate the law. In the seventh, the reference card had changed its label to "not hallucinated" after we narrowed the rule on statements of the form "case law holds" on 23 October.

Search behaviour. The judges differed in how much they used the search they had. On the extension set Gemini 4 Argon ran on average 12.2 searches per card and Sol 6.1 11.2; both searched on every card. Gemini 3.1 Pro ran 4.3 and did not search at all on 25 of 80 cards; Gemini 3.8 Flash ran 4.7 and did not search on 9. Gemini 4 Argon, which searched on every card, had the highest agreement with the reference, and Gemini 3.8 Flash the lowest. The relation is not strict. Sol 6.1 searched on every card and Gemini 3.1 Pro skipped search on 25 cards, yet both reached κ 0.80, with errors in opposite directions. Gemini 3.1 Pro also searched slightly less than Gemini 3.8 Flash but agreed with the reference more often. The number of searches alone does not explain the difference. A second observation concerns how the judges reported what they had not checked. The instruction asked each judge to list the citations it could not check. Sol 6.1 listed on average 0.4 per card; Gemini 4 Argon, Gemini 3.1 Pro and Gemini 3.8 Flash listed almost none (0.01 or fewer). A judge that reports every citation as checked, after zero searches, gives a verdict that cannot be distinguished from a verified one.

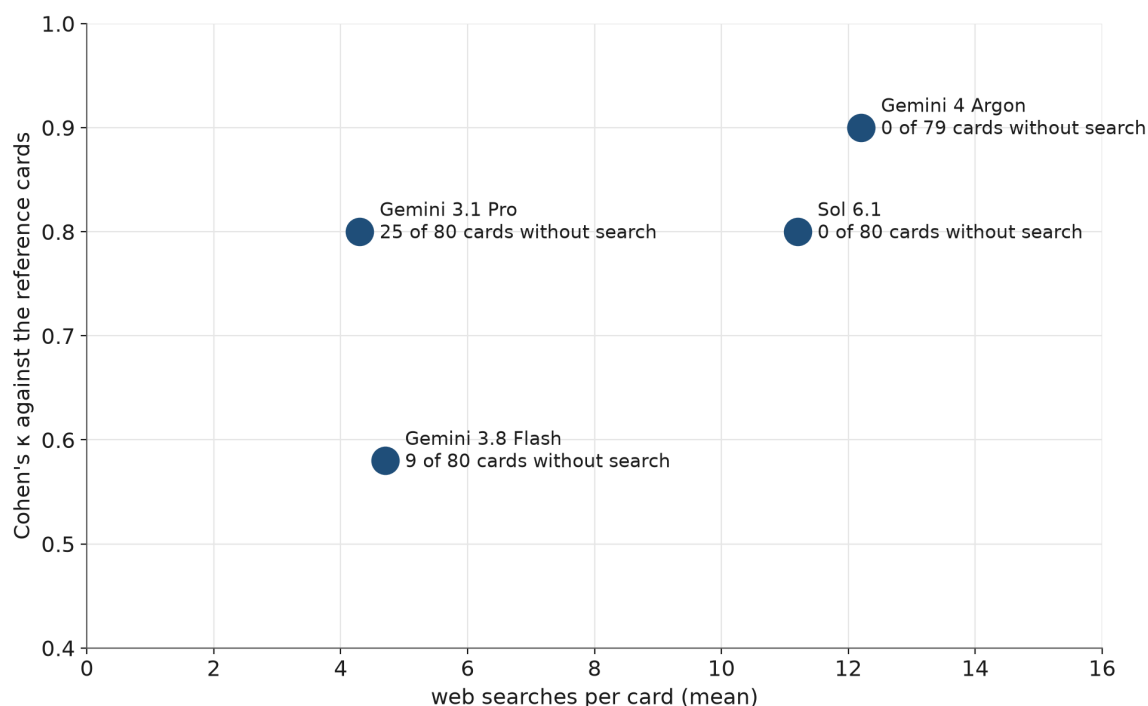


Figure 7. Mean number of web searches per card against κ on the extension set. The judge that searched most agreed most; two judges with different search behaviour reached the same κ with errors in opposite directions.

What the judges missed. Almost every error missed by two or more judges belongs to one of three groups: a provision quoted in wording that was no longer in force on the date of the question (class W); a real case number attached to a thesis that the judgment does not contain (class T); and a sentence of the form "the Supreme Court consistently holds that..." with no case cited. **None of these errors can be detected by reading the answer.** Detection requires the official text for the legal-state date, the full judgment, or the line of case law. In our runs, web search returned these unreliably: it ranked the current consolidated text above earlier versions, it returned summaries and commentaries rather than judgments, and it did not show that a judgment was absent. A judge that does not obtain the source falls back on the same memory as the model that wrote the answer. Two Gemini judges missed more hallucinations in answers from Google models than in answers from other providers: Over all ten questions, Gemini 3.8 Flash missed 13 of the 29 hallucinated answers from Google models (45 %), against 6 of 49 from Anthropic (12 %) and 2 of 13 from OpenAI (15 %); Gemini 3.1 Pro 7 of 29, 6 of 48 and 1 of 13. Gemini 4 Argon showed no clear difference (2 of 24, 1 of 44, 1 of 12; it has fewer scored cards). Google answers in this set contain many class W errors, which all judges detect less well, so part of the difference may come from the type of error rather than from the model family; with 12 to 49 cards per provider we cannot separate the two causes.

Panels. The panel rules we tested did not solve this. Combining the judges by majority vote, by "Argon or Gemini 3.1 Pro", or by "any judge reports it" gave κ 0.90, 0.92 and 0.82 against 0.90 for Gemini 4 Argon alone (Appendix D.4), on the same data used to choose the combination. The remaining errors were not disagreements between judges. They were facts that no judge had checked. A recent study of panels of LLM judges reports the same pattern: nine judges from seven model families provided about two independent votes' worth of information, and the best single judge matched the panel [13]. Our data add a likely cause of the correlated errors: facts that none of the judges verified in the source.

Conclusion. The panel rules we tested changed κ by at most 0.02 on the data used to select them; the differences between single judges were many times larger. The judge that searched most agreed with the reference most, but two judges with different search behaviour reached the same κ , and no judge reported reliably which citations it had not checked. At this stage we recommend improving the verification performed by a single

judge before adding judges to a panel. Every tool call of a judge should be logged: the search query, the source returned, and the version and date of that source. Without such a log a reader cannot check whether a verdict marked as verified was in fact verified; the statistics in this section exist only because the calls were logged. We did not test repeated runs of the same judge, so we make no claim about them.

¹ Judges: Gemini 4 Argon (`gemini-4-argon`, reasoning level "high") and Gemini 3.1 Pro (`gemini-3.1-pro-preview`) and Gemini 3.8 Flash (`gemini-3.8-flash`), all via Vertex AI (Google Cloud, region "global"), temperature 0, with Google Search grounding enabled; Sol 6.1 (`gpt-6.1-sol`) and Astra 6 (`gpt-6-astra`) via the Codex CLI with live web search, reasoning effort "medium", read-only sandbox. Astra 6 scored the pilot set and only 6 cards of the extension set because of account usage limits, and is therefore not reported for the extension set. All judges received the same prompt text. For the Gemini judges we logged the search queries and the sources returned (up to 20 queries and 25 sources per card); for the judges run through the Codex CLI we logged only the number of searches. Judge runs took place between 13 and 20 October 2026; the specification version is recorded in each output file.

9.2 The verification layer

We therefore built the check that the judges lacked. We call it the **verification layer**. It runs before the judge reads the answer, it is independent of the judge, and it has two steps.

- **Citation check** (code only, no model). Code extracts every citation from the answer. For each provision it retrieves the official text as in force on the legal-state date of the question, and the wording on earlier dates, to detect stale wording. For each judgment it looks up the case number in the official database and records the court, the date and the full text. It compares every quoted passage with the source verbatim.
- **Support check** (a model with restricted input). A model receives only the claim that the answer attributes to one source and the text of that source. It decides whether the source supports the claim. It may return "supports" only together with a verbatim quote from the source, and the code then confirms that the quote exists in the source.

The output of the verification layer is the **verification report**. For every citation it gives one of three things: "✓" when every check passed, an error class (F, T, W, S), or a verification status ("source text unavailable", "wording on the legal-state date not established", "ambiguous case number", "act not identified"), together with the official wording, its version and the date it applies to. Statuses are reported separately and are never counted as model errors. A failed check is not evidence that the model was wrong, and the two outcomes must not share a label.

Two results follow. First, on the two questions used to develop the verification layer, the citation check detected every F, W and S error in the reference cards. On the eight questions it had not been developed on (P01, P04, P05, P06, P07, P08, P12, P13; 128 answers; below: the test questions), the results were weaker. The citation check alone detected 3 of 6 F errors and 15 of 44 W errors. The remaining F errors were found by the tool under a neighbouring unit number and were therefore reported as T, and most W errors were outside its reach because the wording on the legal-state date could not be established. For the verification layer as a whole, sensitivity to the error list from the reference cards was about 57%, and false alarms on citations accepted by the cards rose from 1% to 8%. Most of this loss came from input processing, not from judgment: a citation parser that did not recognise a letter superscript in a Polish article number (art. 131^{1a}); a default statute taken from the answer key where the answer named a different act; two judgments sharing one case number (the same number used by different courts). After these were corrected, without any change to the prompts, false alarms fell to 1.8% and sensitivity rose to 59%. Adding two rules from the specification to the prompt of the support check (on the persons to whom a provision applies, and on judgments with the same case number) raised sensitivity to 64% (50% with the exact class) at the same rate of false alarms.

Table 6. The verification layer on the eight test questions (128 answers; 147 citation errors and 665 accepted citations in the reference cards).

Measure	23 Oct, frozen prompts	24 Oct, input fixes	24 Oct, two rules added to the prompt
sensitivity, any class	84 (57%)	87 (59%)	94 (64%)
sensitivity, same class	n/a	66 (45%)	74 (50%)
class T, any class (of 83)	39 (47%)	49 (59%)	50 (60%)
class W, same / any class (of 44)	n/a / 25	17 / 28	24 / 32
false alarms on the 665 accepted citations	53 (8.0%) from the support check, 27 (4.1%) from the citation check	1.8%	1.8%
cost of the run (Vertex AI)	USD 8.09	USD 7.87	USD 7.09

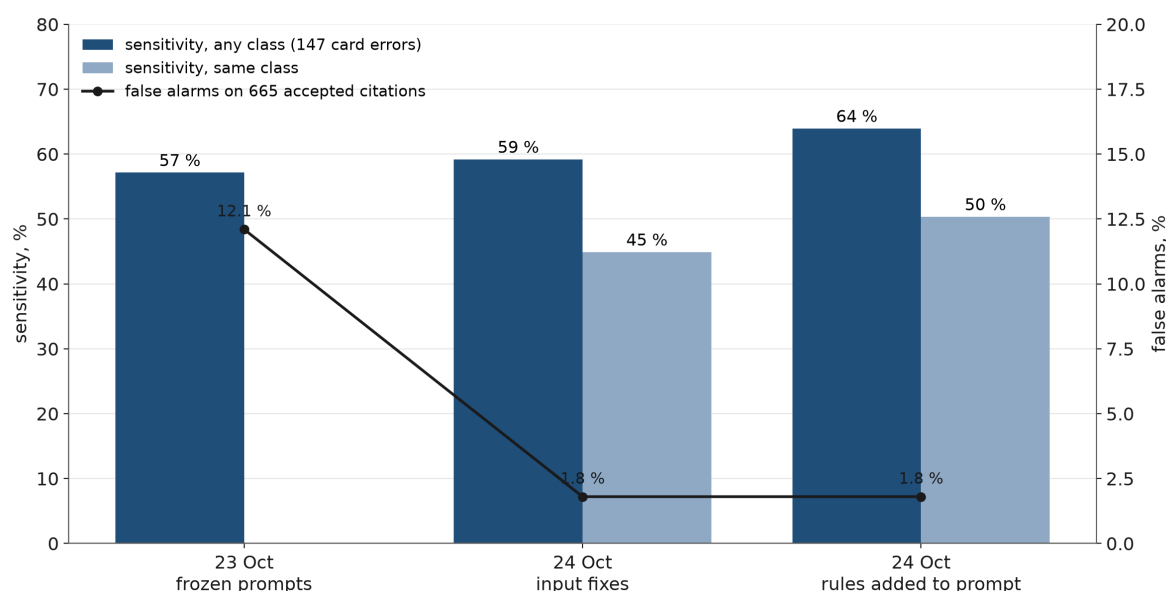


Figure 8. Sensitivity to the errors recorded in the reference cards and false alarms on accepted citations, across the three runs. The fall in false alarms came from fixes to input processing, not from the prompts. Second, class W can be measured only where the official source allows the wording to be reconstructed for a given date. For the Code of Civil Procedure, which has dozens of amendments between consolidated texts, the procedure, which relies on the consolidated texts published in ISAP, returned "wording on the legal-state date not established" for 931 of 941 citations of the Code in the 128 answers. Establishing the wording would require applying the amending acts one by one to the last consolidated text, which the tool does not do.

This points to a general observation. Many of the problems we met did not come from the models' lack of legal understanding. They came from technical issues: retrieving the correct wording of a provision from the official source for a given date, establishing which of several judgments with the same case number, issued by different courts, the answer refers to, obtaining a complete text of a judgment (one database held a version of a Supreme Court judgment about half the length of the official text, without the passage that is usually cited from it), or parsing a citation with a letter superscript. Research on legal benchmarks has so far concentrated on the criteria: how to define "hallucination", what to count, how to weight errors. Our experience suggests that the technical procedure of running a benchmark deserves the same attention. A criterion that cannot be applied, because the source cannot be retrieved in the required version, measures nothing. The verification coverage of

a benchmark, that is, how many citations were actually checked and against which version of the source, should be reported next to its scores.

Conclusion. A verification layer built from citation-checking code, a model-based support check and official sources detected all 7 non-verbatim quotations on the two development questions, but its detection of non-existent units, stale wording and wrong case designations is limited by the official sources it can retrieve and by the quality of its citation parsing. Four faults in input processing (superscript letters in unit numbers, the default act, the same case number in different courts, "in conjunction with" chains) accounted for 94 false alarms in the first run on the test questions and 7 after the fixes. The quality of the sources and of the citation parsing sets the upper limit of what any judge can verify. These technical components need the same design effort as the evaluation criteria.

9.3 A report may add an error; it may not remove one

The next step was to give the verification report to the judge. We tested two judges, Gemini 4 Argon and Sol 6.1, on the eight test questions, that is, the questions on which the verification layer had not been developed, with the input fixes and the two prompt rules of §9.2 already applied, so the results after those changes are exploratory rather than a test on an untouched set (§9.2). The instruction variants were first compared on four of them (P05, P06, P07, P13) and the chosen variant was then run on all eight. In the first variant the report was binding: the judge was instructed to base its hallucination verdict on the classes in the report. On the four questions, agreement with the reference cards decreased for both judges (Argon κ 0.96 $\rightarrow$ 0.85 on 54 cards, Sol 6.1 0.88 $\rightarrow$ 0.83 on 61 cards; the baselines differ between comparisons because each is computed on the cards the variant covered). The mechanism was visible on individual cards. A "✓" in the report meant only that the provision exists, has the stated wording, and that the quotation is verbatim. The judges treated it as a confirmation of the whole citation and stopped checking whether the answer described, for example, the circle of addressees of a provision as it was before an amendment. In the second variant the report was asymmetric: a class F, T, W or S from the report is binding; a "✓" is not, and the judge must still compare what the answer attributes to the source with the wording given in the report. With this instruction, on the four questions, Sol 6.1 exceeded its own baseline (κ 0.89 $\rightarrow$ 0.96; false alarms 3 $\rightarrow$ 1; no new misses), while Argon fell from 0.96 to 0.88 (misses 1 $\rightarrow$ 3). On all eight test questions the pattern held. **Sol 6.1 rose from κ 0.84 without the report to 0.97 with it:** seven of its eight false alarms disappeared and no misses were added (119 cards). **Argon fell from 0.95 to 0.88** (115 cards). Its three additional misses were all errors of class W that the verification report had shown as correct. In one, the wording on the legal-state date could not be established; in one, the report had been produced before the rule on the circle of addressees was added to the support check; in one, the answer assumed the wrong year for a statutory threshold, which the check does not test.

We then ran a third variant on all 128 cards to see how the judges react to an explicit status in place of a confirmation mark. In it the report shows "✓" only when the wording on the legal-state date was confirmed; otherwise it shows the status "wording on the date not established" and tells the judge to assess the citation itself. On the 119 cards scored under both the second and the third variant (the first variant covered four questions only), Sol 6.1 reached κ 0.91 with the third variant against 0.97 with the second (false alarms 8 $\rightarrow$ 1 $\rightarrow$ 2, misses 1 $\rightarrow$ 1 $\rightarrow$ 3); Argon stayed at 0.88 on its 115 shared cards. Sol 6.1's two new misses were class W errors which the same judge had detected under the second variant. In both cards the report now showed the status instead of "✓"; the judge assessed the wording itself, accepted it as current, and in one case treated a court fee quoted at a repealed amount as an omission rather than as class W. The status did not prompt the judge to check the source. It removed a cue which the judge had used under the second variant, and the judge then relied on its own reading of the answer. The 95 % intervals, obtained by bootstrap over questions, overlap for the second and third variants (0.92–1.00 and 0.85–0.98 for Sol 6.1) and, at the edge, for Sol 6.1 without and with the auxiliary report (0.70–0.96 and 0.92–1.00); the point estimates differ in a consistent direction, but with eight questions neither the size of the difference nor its direction beyond this sample is resolved (Appendix D. 5).

Table 7. Judge without and with the verification report, by variant of the instruction (κ on the label "hallucinated"; same cards within a row; intervals from Appendix D.5).

Judge	Variant	n	κ without report	κ with report	95 % interval with report	False alarms without $\rightarrow$ with	Misses without $\rightarrow$ with
Gemini 4 Argon	v1 binding	54	0.96	0.85	0.76–0.95	0 $\rightarrow$ 0	1 $\rightarrow$ 4
Gemini 4 Argon	v2 auxiliary	115	0.95	0.88	0.86–0.91	0 $\rightarrow$ 1	3 $\rightarrow$ 6
Gemini 4 Argon	v3 status	116	0.95	0.88	0.86–0.91	0 $\rightarrow$ 2	3 $\rightarrow$ 5
Gemini 4 Argon, same cards	v2 / v3	115	0.95	0.88 / 0.88	0.86–0.91 / 0.85–0.91	0 $\rightarrow$ 1 / 2	3 $\rightarrow$ 6 / 5
Sol 6.1	v1 binding	61	0.88	0.83	0.75–0.93	3 $\rightarrow$ 0	0 $\rightarrow$ 5
Sol 6.1	v2 auxiliary	119	0.84	0.97	0.92–1.00	8 $\rightarrow$ 1	1 $\rightarrow$ 1
Sol 6.1	v3 status	128	0.80	0.89	0.82–0.95	11 $\rightarrow$ 4	1 $\rightarrow$ 3
Sol 6.1, same cards	v2 / v3	119	0.84	0.97 / 0.91	0.92–1.00 / 0.85–0.98	8 $\rightarrow$ 1 / 2	1 $\rightarrow$ 1 / 3

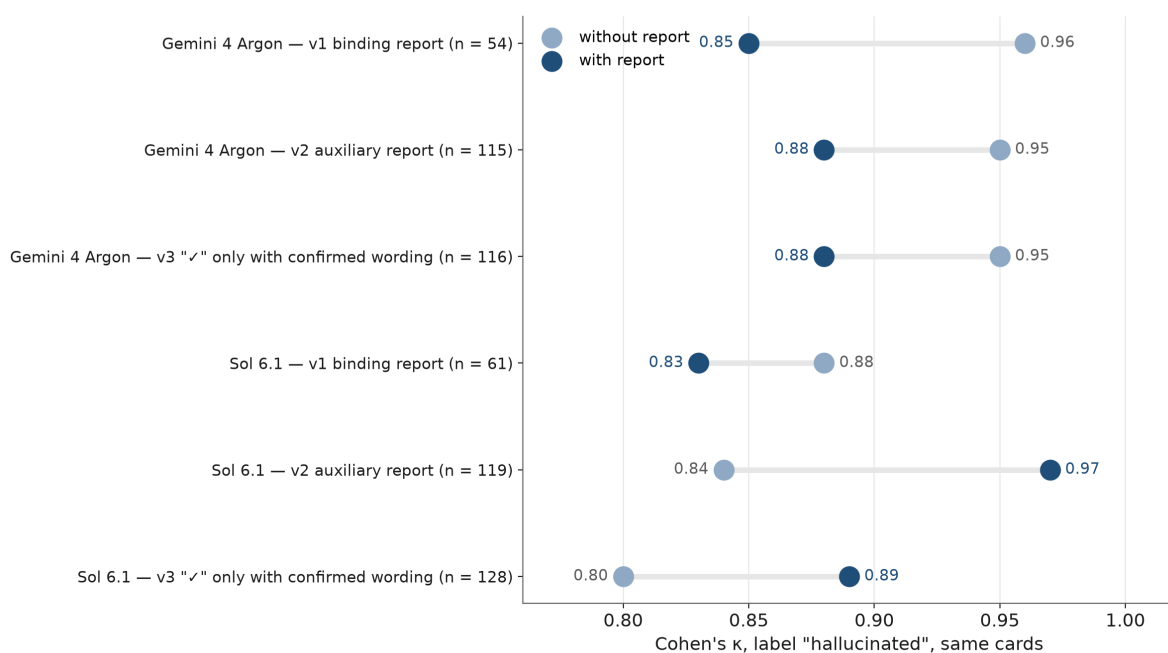


Figure 9. Agreement with the reference cards without the report (grey) and with it (blue), on the same cards, for three variants of the instruction. We state this as a working hypothesis: the difference between the variants rests on two cards of one judge, and the mechanism should be tested on a larger sample and on more than two judges before it is treated as a result.

This agrees with work on agents, which adopt corrupted tool outputs at high rates, up to 68% for web search [14]. It also needs one qualification. When a supplied reference contradicts what a judge already believes, judges tend to ignore it [15]. The effect we observed concerns a report that agrees with what the judge would accept anyway: a "✓" next to a citation for which the judge recorded no further check of its own. We propose the following rule for further testing: an error class from the automated verification binds the judge, while a

"✓" does not relieve the judge of its own check. The rule does not freeze the judge's earlier verdicts: Sol 6.1 withdrew seven of its own false alarms after seeing the official wording in the report. A check that confirms existence and wording does not confirm meaning, and in our runs judges instructed to treat it as confirmation accepted such citations without further checking.

Conclusion. **The agreement of a judge varied with the instruction that came with the verification report.** A binding report lowered the agreement of both judges. A report that can only add errors improved the stricter judge, but the other judge still lost detections where the report showed a citation as correct although the check had been incomplete. Showing a status instead of a confirmation mark changed nothing for that judge and was associated with two additional misses for the stricter judge. We recommend that automated checks confirm errors, that a citation is never shown as correct unless every element of the check was completed, and that the absence of errors is left to the judge. The judge must therefore be instructed explicitly to check every citation in the answer against the official source itself, including the citations that the verification report marks as correct.

9.4 Cost

The citation check has no cost other than the network requests to the official sources. The support check costs about USD 0.006–0.008 per citation. A judge card produced with the verification report costs about USD 0.10–0.13 in tokens for a reasoning model on Vertex AI. This is two to three times the cost of a judge card without the report, because the prompt grows to about 19,000 tokens. Web searches by Gemini judges are billed separately above a monthly free quota; at the list price they add about USD 0.05–0.07 per card with the report and about USD 0.20 per card for Gemini 4 Argon without it. For a benchmark of 160 answers, one full pass costs tens of dollars. The whole study, including the answers, the judges and all runs used to develop the verification layer, cost at least USD 170 in model fees, with searches at the list price; runs through subscriptions are not included. For comparison, preparing the answer keys and closing the 160 reference cards took the author two to three weeks of work; the reference cards remain the standard against which all of these figures are measured.

Table 8. Model cost of the study (USD, price lists of 18 October 2026; lower bound).

Part	Token cost	Web searches at list price	Total
160 evaluated answers	≥ 15.78	included	≥ 15.78
model judges, Vertex AI (Codex runs through a subscription not included)	≥ 13.82	55.02	≥ 68.85
verification layer, all development and test runs	66.24	19.82	86.07
Total	≥ 95.85	74.84	≥ 170.70

Sums are computed on unrounded amounts, so rows may differ from the sum of their displayed parts by one cent.

Conclusion. **Model cost is not the limiting factor; the lawyer's time on the reference cards is.** A judge pass with a verification report costs tens of dollars for a benchmark of this size, excluding human review. Rescoring a fixed set of answers is cheap to repeat; a new model gives new answers, which need new reference cards before its score means anything.

9.5 Answer

For the binary verdict on a civil-law legal question, a reasoning model with web search and a verification report used as auxiliary evidence reached κ 0.97 against the reference cards on eight test questions (Sol 6.1). The same report lowered the agreement of a second judge (Gemini 4 Argon, κ 0.95 $\rightarrow$ 0.88), which deferred to citations shown as correct without a complete check. When those marks were replaced with an explicit status, the judges did not treat "not established" as a reason to check the source. **The source of the change differs**

between the judges. Sol 6.1 improved because seven of its eight false alarms disappeared, not because it detected new errors: its one miss is the same card before and after. Argon lost agreement because its misses rose from three to six. Our working hypothesis is that both judges followed the report: where the judge's own verdict was a false alarm which the report did not confirm, the judge withdrew it; where the report showed as correct a citation whose check had been incomplete, the judge accepted it. One hypothesis is that the effect of the report depends on the judge's own error profile: it helps a judge whose errors are false alarms and harms a judge whose errors are misses. The hypothesis rests on a small number of cards from two judges, without repeated runs, and needs a larger sample. Without the report, both judges were accurate on reasoning errors and less accurate on the errors that make legal hallucination harmful: stale wording, a real case number carrying a thesis from another case, a line of case law that does not exist. Two caveats remain. The reference cards were reviewed and closed by one lawyer, the author. The agreement figures on the pilot set are inflated by the order in which cards and judge outputs were produced. Both caveats call for repeating the measurement on questions that neither the judges nor the verification layer have processed before. That is the next step.

10. The same conditions in evaluation and in knowledge representation: OKF-Legal

A benchmark measures how often a model's answer is not correctly grounded. It does not specify how a legal knowledge system should be built so that the answer can be correctly grounded. These are two sides of one problem. On both sides the same conditions appear.

Taranukhin and Shwartz [4] start from evaluation. A legal claim is warranted when the authority behind it exists, applies in the jurisdiction, is current for the date of analysis, has the legal status the system attributes to it, and supports the proposition. They add a response policy (answer, narrow, ask, warn, abstain) and propose jurisdiction-specific ontologies of authority as future work.

We arrived at a similar list independently, starting from knowledge representation. OKF-Legal [11] is a legal profile over Google's Open Knowledge Format [12]. It was derived from two working systems built in our firm: a case-law corpus and a contract-drafting skill. It was not designed from the warrant conditions; it was developed in response to production incidents. Its fields and consumption rules nevertheless correspond closely to the warrant conditions:

Table 9. Warrant conditions, OKF-Legal fields and our error classes.

Warrant condition [4]	OKF-Legal [11]	Error class in our cards
the authority exists	resolver to the official source (ELI, EUR-Lex), rule R-RESOLVER	F
it supports the proposition	verbatim anchor quote, rule R-CYTAT	T
it is current for the date of analysis	obowiazuje-od/do (in force from/to), rule R-CZAS	W
— (not distinguished)	stосуje-sie-od (applies from): current text ≠ norm already binding	W
it has the status represented (dicta, dissent, lower court)	fields tryb (ratio / obiter / referred), instancja-zrodlowa, zrodlo-wypowiedzi (who said it)	T or S
a source on the right topic that does not support the proposition	rule R-NIEWIEM: the nearest search result is not evidence	T
response policy: warn, abstain	rules R-BRAK and R-NIEWIEM: signal uncertainty instead of assuming	B2 (calibration)
jurisdiction-specific ontology of authority (proposed)	vocabulary of concept types for Polish law (§3 of the profile [11])	—

Two observations follow. First, one row exists only on the representation side: the distinction between the date from which a text is current and the date from which the norm applies. For EU acts with staggered application dates, such as the AI Act, an answer can quote the current text correctly and still be wrong about whether the obligation already applies. Our error classes would need this distinction as well. Second, several errors found in the benchmark are errors which these fields were introduced to prevent. One model attributed to the Supreme Court a view which in the source was the question of the referring court; in OKF-Legal that statement would carry `tryb: referowana` and the name of its actual author. Another model, with web search, opened the correct appellate judgment and then wrote the case number of a different judgment; a verbatim anchor quote tied to the source identifier would make this mismatch detectable by a consistency check.

This comparison has limits. OKF-Legal is the author's own proposal, and I am not a neutral assessor of it. Its retrieval test (30 questions, a model judge blind to the method) is small and model-scored. A format does not ground an answer by itself; it only makes grounding checkable. What the convergence shows is limited to the two lists: two groups, one starting from evaluation in common law and one from building systems in Polish civil law, produced lists that largely overlap. For practice this means that the benchmark criteria and the design of the knowledge base can share one vocabulary, and that the verification report of a model judge (§9.2) can apply the same rules that a legal assistant should follow when it answers.

Conclusion. **Evaluation and knowledge representation meet the same three conditions:** the source must exist, it must be current on the legal-state date, and it must support the claim attributed to it. A benchmark that scores these conditions and a knowledge format that records them can use one vocabulary. Whether a format in use reduces the errors measured here is an empirical question, which this paper does not answer.

11. What legal AI is actually good for right now

In my assessment, the value of "legal AI" today is not in the model's memory of the law. It is in two other things: access to an up-to-date knowledge base (current wording of statutes, current case law) and reasoning sufficient to understand what a judgment says. Magesh et al. make a related point: retrieval itself requires legal reasoning, because the model has to identify the issue first and only then search for the sources. Identifying which provisions to search for requires a proper legal framework. Our numbers point the same way, with the limits noted in §6.2: in the pilot, 41 of 50 answers from memory were labelled hallucinated against 2 of 15 from systems with search, though the two groups were different systems; in the paired W0/W1 set the same eight models had the label in 51 of 80 answers without tools and 40 of 80 with search.

12. Conclusion

The edge cases listed here are only the ones we can distinguish so far. Further work on what "correctly grounded" means in practice is needed: a validation sample scored by human lawyers, a larger question set across several branches of law, per-citation rates reported alongside the binary verdict, and an explicit weighting of a wrong case identifier against incorrect content attributed to a provision in civil-law systems. The results in §9 add one item to this list: the accuracy of the model judges and of the verification report should be measured again on questions that neither the judges nor the verification layer have processed before. Comparable measurements in other jurisdictions would show which of these observations hold beyond Polish law.

References

1. Magesh V., Surani F., Dahl M., Suzgun M., Manning C. D., Ho D. E., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, Journal of Empirical Legal Studies 22(2), 2025, pp. 216–242, <https://doi.org/10.1111/jels.12413>; preprint arXiv:2405.20362 (2024). The paper evaluates RAG tools (Lexis+ AI, Westlaw AI-AR, Ask Practical Law AI, GPT-4) on US law.

2. Dahl M., Magesh V., Suzgun M., Ho D. E., *Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*, *Journal of Legal Analysis* 16(1), 2024, pp. 64–93, <https://doi.org/10.1093/jla/laae003>.
3. Hu Y., Gan L., Xiao W., Kuang K., Wu F., *Fine-tuning Large Language Models for Improving Factuality in Legal Question Answering*, *Proceedings of COLING 2025*, pp. 4410–4427, <https://aclanthology.org/2025.coling-main.298/>; arXiv:2501.06521 (LegalHalBench, metric NHSR).
4. Taranukhin M., Shwartz V., *Legal LLM Hallucination Should Be Evaluated as Failure of Legal Warrant*, *AI4Law @ ICML 2026*, arXiv:2609.17546.
5. Choi J., Sakaguchi K., Yamada H., *Expert Evaluation of LLM's Open-Ended Legal Reasoning on the Japanese Bar Exam Writing Task*, accepted to ICAIL 2026, arXiv:2604.23730 (April 2026): 295 of 441 precedent mentions vs 381 of 2,575 statute mentions hallucinated; one law-school-graduate annotator.
6. Figueiredo A. et al., *Grounded in Law: A Multi-Stage Anti-Hallucination Pipeline for Legal RAG Systems in Brazilian Portuguese*, *Proceedings of PROPOR 2026*, Vol. 2, pp. 30–34, <https://aclanthology.org/2026.propor-2.9/>: 43,175 references, legislation resolved 81.7%, jurisprudence 47.1%.
7. Piotrowski A., Koman A., *LexAlpha Lab · Benchmark 02 · pytania prawne* (Polish; working report: scoring specification, question set, model judges, verification layer, cost), LexAlpha Lab, working version of 25 October 2026, available from the authors on request; to be published with the question set.
8. Supreme Administrative Court of Poland, order of 23 June 2026, I FZ 104/26; judgment of 31 August 2026, I OSK 590/26 (as reported by [infor.pl](https://www.infor.pl/prawo/nawosci-prawne/7623766,ai-w-pismach-procesowych-do-sadu-halucynacje-postanowienie-nsa-2026.html), 17 July 2026, <https://www.infor.pl/prawo/nawosci-prawne/7623766,ai-w-pismach-procesowych-do-sadu-halucynacje-postanowienie-nsa-2026.html>, and [prawo.pl](https://www.prawo.pl/prawnicy-sady/nsa-radca-prawny-przywolal-orzeczenia-zmyslone-przez-ai,1553132.html), 22 September 2026, <https://www.prawo.pl/prawnicy-sady/nsa-radca-prawny-przywolal-orzeczenia-zmyslone-przez-ai,1553132.html>).
9. Karp M., Kubaszewska A., Król M., Król R., Smywiński-Pohl A., Szymański M., Wydmański W., *LLM-as-a-Judge is Bad, Based on AI Attempting the Exam Qualifying for the Member of the Polish National Board of Appeal*, arXiv:2511.04205 (November 2025).
10. Solicitors Regulation Authority, *Misuse of AI* (warning notice), 17 August 2026, <https://www.sra.org.uk/solicitors/guidance/misuse-ai/>; CCBE, *CCBE guide on the use of generative AI by lawyers*, 2 October 2025, https://www.ccbe.eu/fileadmin/speciality_distribution/public/documents/IT_LAW/ITL_Guides_recommendations/EN_ITL_20251002_CCBE-guide-on-the-use-of-the-use-of-generative-AI-for-lawyers.pdf.
11. Piotrowski A. (KTZR), *OKF-Legal: a legal profile over Open Knowledge Format*, version 1.0, August 2026, Apache 2.0, <https://github.com/apiotrowski-afk/okf-legal> (author's own proposal; English summary in [SPEC.en.md](https://github.com/apiotrowski-afk/okf-legal/blob/main/SPEC.en.md)).
12. Google Cloud, *Open Knowledge Format (OKF)*, v0.2, <https://github.com/GoogleCloudPlatform/knowledge-catalog/tree/main/okf>.
13. Kohli G., *Nine Judges, Two Effective Votes: Correlated Errors Undermine LLM Evaluation Panels*, preprint, 2026, arXiv:2605.29800.
14. Yang H., Song W., Kim T., Song J., Park S., Jo Y., *Agents' Overreliance on Unreliable Tools*, preprint, 2026, arXiv:2609.05587.
15. Lee D., Hwang Y., Kang T., Lee M., Chae Y., Jung K., *Judging Against the Reference: Uncovering Knowledge-Driven Failures in LLM-Judges on QA Evaluation*, preprint, 2026, arXiv:2601.07506.
16. Butler A.-R., Butler U., *Legal RAG Bench: an end-to-end benchmark for legal RAG*, preprint, 2026, arXiv:2603.01710.
17. Bang Y., Ji Z., Schelten A., Hartshorn A., Fowler T., Zhang C., Cancedda N., Fung P., *HalluLens: LLM Hallucination Benchmark*, in: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2025, pp. 24128–24156.

Appendix A. Systems, data sets and model judges

A.1 Pilot run (r3, 10 October 2026)

All answers collected on 10 October 2026; legal-state date: 1 October 2026; 5 questions per system.

Models without tools (W0: no search, no tools, no access to statutes)

Model	Identifier	Interface	Settings
Astra 6	gpt-6-astra	Codex CLI	reasoning effort medium
Sol 6.1	gpt-6.1-sol	Codex CLI	reasoning effort medium
Terra 5.6	gpt-5.6-terra	Codex CLI	reasoning effort medium
Fable 5.1	claude-fable-5-1	Claude Code CLI	default temperature
Opus 5.5	opus (alias)	Claude Code CLI	default temperature
Sonnet 5.5	claude-sonnet-5-5	Claude Code CLI	default temperature
Haiku 5.5	claude-haiku-5-5	Claude Code CLI	default temperature
Gemini 3.1 Pro	gemini-3.1-pro-preview	Vertex AI	temperature 0.1
Gemini 2.5 Pro	gemini-2.5-pro	Vertex AI	temperature 0.1
Gemini 3 Flash	gemini-3-flash-preview	Vertex AI	temperature 0.1

Systems with web search

System	Interface	Notes
ChatGPT, temporary chat	chatgpt.com	no account memory; model not disclosed by the app
ChatGPT, new chat	chatgpt.com	model not disclosed by the app
Sol 6.1 with search	Codex CLI	gpt-6.1-sol, reasoning effort medium

A.2 Extension run (18 October 2026): the same models with and without search

On 17 October 2026 the benchmark design was changed so that every model answers every question twice: without tools (W0) and with web search (W1). Eight models from three providers took part; the two ChatGPT application variants and Gemini 2.5 Pro and Gemini 3 Flash from the pilot were not continued. The same system instruction was used as in the pilot. Answers to the five new questions were collected on 18 October 2026; the W1 answers to the five pilot questions were collected on the same day.

Model	Identifier	Interface	W0	W1 (search tool)
Sol 6.1	gpt-6.1-sol	Codex CLI, reasoning effort medium	yes	web search (Codex)
Astra 6	gpt-6-astra	Codex CLI, reasoning effort medium	yes	web search (Codex)
Terra 5.6	gpt-5.6-terra	Codex CLI, reasoning effort medium	yes	web search (Codex)
Opus 5.5	claude-opus-5-5	Claude Code CLI	yes	WebSearch, WebFetch (page summary by Haiku 5.5)
Sonnet 5.5	claude-sonnet-5-5	Claude Code CLI	yes	WebSearch, WebFetch
Haiku 5.5	claude-haiku-5-5	Claude Code CLI	yes	WebSearch, WebFetch
Gemini 4 Argon	gemini-4-argon	Vertex AI, reasoning high, temperature 0.1	yes	Grounding with Google Search, URL context
Gemini 3.8 Flash	gemini-3.8-flash	Vertex AI, temperature 0.1	yes	Grounding with Google Search, URL context

Two features of the W1 condition limit what can be measured. Claude Code does not record the number of searches, and its WebFetch tool returns a summary of the page written by a smaller model rather than the page text. The Vertex

AI runs of 18 October did not record the tokens of page content read through URL context. Both facts make the W1 cost a lower bound and leave the origin of some W1 errors undetermined (Appendix F).

A.3 Data sets used in this paper

Set	Questions	Systems	Answers	Cards closed	Used in
Pilot run r3	P05, P06, P07, P12, P13	13 (10 W0, 3 with search)	65	after the judges' outputs were seen	§6; §9.1 (agreement on the pilot set)
Extension set	P01, P04, P08, P09, P10	8 models × W0/W1 = 16	80	before any judge ran	§9.1
Test set of the verification layer	P01, P04, P05, P06, P07, P08, P12, P13	16	128	mixed (see above)	§9.2, §9.3
Full W0/W1 set	10 questions	16	160	mixed	Appendix F, Appendix G; 91 answers hallucinated (51 W0, 40 W1), 46 correctly grounded (19 W0, 27 W1)

The four sets overlap. The 65 pilot answers and the 160 W0/W1 answers share the 30 W0 answers to the pilot questions from the six pilot models that continued (Sol 6.1, Astra 6, Terra 5.6, Opus 5.5, Sonnet 5.5, Haiku 5.5), re-scored under the current specification; Gemini 4 Argon and Gemini 3.8 Flash answered the pilot questions without tools on 13 and 17 October 2026; the pilot figures in §6 are those of the specification in force on 17 October 2026 and are not restated here. Questions P02, P03 and P11 have answer keys from the pilot period but are not in any set used in this paper: P02 was excluded as a computation task; P03 and P11 (consumer credit) were not taken into the extension and will be included in the next run.

A.4 Model judges

Judge	Identifier	Interface	Search	Settings	Specification
Gemini 4 Argon	gemini-4-argon	Vertex AI (global)	Grounding with Google Search	temperature 0, reasoning high	1.0.1
Gemini 3.1 Pro	gemini-3.1-pro-preview	Vertex AI (global)	Grounding with Google Search	temperature 0	1.0.1
Gemini 3.8 Flash	gemini-3.8-flash	Vertex AI (global)	Grounding with Google Search	temperature 0	1.0.1
Sol 6.1	gpt-6.1-sol	Codex CLI	web search (live)	reasoning effort medium	1.0.1
Astra 6	gpt-6-astra	Codex CLI	web search (live)	reasoning effort medium	1.0.1; 49 cards only (account limits)

Every judge received, in one prompt, the scoring rules (blocks A and B with the boundary examples and recorded decisions), the full answer key of the question, and the text of the answer without the name of the system. The judge returned a structured card with the same fields as the reference card. The judges did not see the reference card or each other's outputs. Each Vertex AI judge call recorded the number of search queries and the queries themselves; for Codex judges the number of web-search events was taken from the CLI event log (queries recorded from 25 October 2026).

Appendix B. Scoring specification

The scoring specification is version 1.0.1 of 11 October 2026, with the boundary examples of 17 October and the decisions recorded until 25 October 2026. The full text is in Polish in the benchmark repository; this appendix restates the parts needed to read the paper.

B.1 Question types

Type	What it measures	How the core (A-R) is scored
closed	hit: yes / no / "depends on X"	agreement with the key's core
computation	a numeric result	agreement with the key's result within a tolerance
verification	whether the system checks a claim put to it or repeats it	correct verdict on each claim
trap (feature, not a category)	whether the system notices a wrong or outdated premise in the question	premise named and corrected
interpretive	required distinctions	distinctions from the rubric
sources	retrieval of case law	A6 and a list of known judgments
checklist	completeness	share of the list's elements

A trap is a property of a question of any type. Six of the ten questions in the W0/W1 set contain one (Appendix C).

B.2 Block A (objective)

#	Criterion	Points
A-R	core of the answer, stated anywhere in the answer	0 / 1 / 2 (2 = full core; 1 = correct direction without the key condition)
A2	required legal bases	0 / 1 (0 if any required basis is missing or wrong)
A3	quotations of provisions	0 / 1, only if the system quotes; otherwise not applicable
A5	disputed status named	0 / 1, only for questions marked as disputed in the key
A6	hallucination	1 / 0 (binary since 11 October 2026); fields: error class, origin, scope
A8	framing: the answer addresses the question asked, without digression	0 / ½ / 1
R	rubric elements of the key	+1 each
K	penalty elements of the key	-1 each (A does not fall below 0)

A % = (points obtained – penalties) / points available. Points available exclude criteria that are not applicable.

B.3 Block B (judgement)

#	Criterion	Points
B1	usefulness: the lawyer knows what to do next	0–2; a factual error in a practical instruction (fee, deadline, authority, form) caps B1 at 1 even when no provision is cited
B2	calibration of confidence against the key's expected calibration (categorical / conditional)	0–1

$\Delta = A \% - B \%$. Block B is reported separately from block A and does not duplicate it: a confident wrong answer to a categorical question keeps B2 = 1 and pays in A-R and A6.

B.4 Error classes

A6 is 0, and the answer carries the label *hallucinated*, when at least one statement about a source of law is inconsistent with the official text on the legal-state date. The class is recorded for every faulty citation and does not change the score. One answer may carry several classes.

Class	Definition	Example from the cards (model, condition, question)
F	the source or the unit does not exist (case number, act, article, paragraph outside the numbering)	Claude Sonnet 5.5, W1, P13: art. 1043 ¹ of the Code of Civil Procedure cited as a basis for a deposit; no such unit exists
T	the source exists but does not contain the attributed content: another norm, another thesis, the opposite of the ruling, a judgment that did not decide the point, a wider or narrower scope, another addressee; also a thesis attributed to "case law" in general and contradicted by a judgment in the key	Gemini 3.8 Flash, W0, P07: a "quotation" of art. 357 § 5 of the Code of Civil Procedure with content the provision does not contain (the official text says only that principal reasons are stated in a way that distinguishes them from a justification). Claude Opus 5.5, W0, P07: "in practice courts reject it or leave it without action", a thesis attributed to court practice without a source and contrary to art. 357 § 1, § 2 ¹ and § 5
W	the provision exists but is quoted or described in the wording of another date than the legal-state date; also a wrong history of amendments	GPT-5.6 Terra, W1, P07 and Gemini 3.8 Flash, W1, P07: court fee of PLN 100 under art. 25b of the Act on Court Costs in Civil Cases, a repealed wording. Claude Haiku 5.5, W0, P01: the threshold of a petty offence against property given as one quarter of the minimum wage, the wording in force until 14 November 2018
S	a real source with a distorted identifier (case number, date, court, publisher, name of the act), with a thesis consistent with the real source	Gemini 3.8 Flash, W1, P13: case number VIII Cz 83/16 instead of VIII Gz 43/16, thesis correct. Claude Sonnet 5.5, W1, P07: a view of the Warsaw Court of Appeal, quoted in the certified question in III CZP 38/20, attributed to the Supreme Court itself

Three consequences of the definition matter for comparison with other benchmarks. First, an omission is not a hallucination: leaving out a condition or an element of a provision is not a false statement; it costs rubric points and the core. Second, an answer without any citation is not a hallucination; it loses the points for legal bases and carries the status "no source verified". Third, A3 and A6 never fall for the same mistake: a quotation that matches no real version of the provision lowers only A3; a quotation that matches a real version of another date is class W.

B.5 Origin of the error and verification status

For every faulty citation the card records where the error came from. The field does not change the score.

Origin	Meaning
memory	the system did not open the source while answering: a run without tools, tools available but unused, or no trace of that act or judgment in the tool log; the content comes from the model's weights
official source misread	the tool trace or an address in the answer shows that the system opened the right official source and still reported it wrongly
secondary source	the system relied on a portal, blog or commentary and repeated its error or shortcut
undetermined	no tool trace and no addresses; cannot be decided (Claude Code W1 runs)

Each card also carries a status of source verification for the answer as a whole: "no tools" (W0), "tools unused" (W1, tools available but not used), "secondary sources only", or "official source used". The status is descriptive; it is not a score.

B.6 Scope of the hallucination: core or peripheral

Since 18 October 2026 every card with A6 = 0 records whether the error concerns the core of the question or a peripheral remark. A6 stays binary and strict regardless of scope.

Scope	Definition
core	the false statement serves the main conclusion of the answer, concerns a required legal basis, or lies in the content of a rubric element
peripheral	the error is in a side remark, analogy, background or digression, while the main conclusion and its bases are correct

Two clarifications were recorded on 25 October 2026 when the field was filled in for the pilot cards. The field describes the place of the error, not its effect on the score: an error in the content of a rubric element is core even when the element was credited (for example, the right fee named but at a repealed amount); a distorted identifier (class S) with a correct thesis is peripheral even when the judgment supports the main conclusion, unless the judgment is a required basis in the key. When an answer has errors of both kinds, the card records core.

B.7 Boundary rules used in this paper

The rules below were each set on a specific card, which then serves the judges as a reference example. The list is limited to the rules that bear on the results reported in §6 and §9.

- 1. **A unit with another unit's content is T** (10 October). If the system names a specific paragraph, section or point and attributes to it content which that unit does not contain, the error is T even when the right norm is in a neighbouring unit of the same article. Reason: a lawyer copies the cited unit into a brief. Exception (22 October): a neighbouring unit that carries the substance of the attributed content is an imprecision, not T.
- 2. **A repealed unit with its former, true content is W, not F** (22 October).
- 3. **A thesis attributed to "case law" without a citation** (B9, narrowed on 23 October): T only when a judgment in the key explicitly rejects the attributed thesis; statements about the "uniformity" of a line of case law are not penalised.
- 4. **Wrong history of amendments is W** (B8, 18 October), in the core when the history is a rubric element.
- 5. **Old editorial wording without a change of the norm lowers only A3; old wording with a substantive change is W** (B10, B10a, 18 October). A changed defined term in a quotation gives A3 = 0 (rule 19).
- 6. **A narrowed or widened circle of addressees or scope of a norm** (24 October): W when it matches the wording of another date; T when it matches no wording; mere silence about a possibility is an omission, not a false statement.
- 7. **An editorial recommendation presented as a statutory requirement is A8 = ½, not T; a definition narrowed against official guidelines is an error of interpretation in A-R, not T** (19 October).
- 8. **Soft law** (Commission guidelines, codes of practice) enters the key as a separate rubric element, never as a penalty (18 October).
- 9. **Provider block**: an answer that refuses on provider-policy grounds is recorded with a separate flag and scored on what it says (18 October); one such answer is in the set.
- 10. **An error inherited from the question** (11 October): if the system takes a wrong provision number from the question and adds content to it, the error is T, as with its own statement; a caveat that "the numbering should be checked" does not remove it.

B.8 "Correctly grounded"

An answer is correctly grounded when A6 = 1 (no statement about a source of law inconsistent with the official text on the legal-state date) and A-R = 2 (full core). The measure is reported alongside the hallucination label; a cost per correctly grounded answer is an auxiliary metric (Appendix G). The definition is deliberately strict in one direction and lenient in another: one wrong unit anywhere in the answer removes the status, but an answer without any citation can hold it if its core is right and it is marked "no source verified".

B.9 Decisions recorded during the study

Date (2026)	Decision	Effect on the cards
17 Oct	Boundary examples A8 1–3 and B1 1 (digression following a wrong thesis; wrong legal construction in an answer to the question asked; wrong type of ruling throughout; control step next to a thesis that risks a missed deadline)	scores changed in 5 cards; labels unchanged
17 Oct	The firm's own systems excluded from the benchmark; W1 condition for all models	124 files archived; 80 new runs

Date (2026)	Decision	Effect on the cards
18 Oct	Scope field (core / peripheral); provider block; soft law in the keys; cost metric; status "tools unused"	new fields in the extension cards
18 Oct	B8: wrong history of amendments = W, core when the history is a rubric element	P01
18 Oct	Keys P09 v4 and P10 v3: conditional core, Supreme Court line as a rubric element, rule B6 repealed	20 cards re-scored
18 Oct	Decisions 1–34 on the extension cards; B10 / B10a (old editorial wording = A3 only; old wording with substantive change = W); rule 19 (changed defined term = A3 0); B11 (B2 under conditional calibration); B12 (line of case law: interpretation under art. 65 of the Civil Code suffices); no double penalty in block B; A8 = 1 for correct side remarks	59 of 80 cards changed
19 Oct	Editorial recommendation as statutory requirement = A8 ½, no T; definition narrowed against guidelines = interpretation error in A-R, no T	5 cards P08, 1 card P09
22 Oct	Neighbouring unit carrying the substance = imprecision, no T; repealed unit with its former true content = W, not F	P10 (2 cards)
23 Oct	B9 narrowed: a statement about case law without a citation is T only when a judgment in the key explicitly rejects it	1 card: A6 0 → 1
24 Oct	Narrowed or widened circle of addressees or scope: W if it matches another date's wording, T if it matches none; silence is an omission	rule also for the verification layer
24 Oct	Verification layer: S on judgments with assessment of each candidate for an ambiguous case number; report to the judge may add an error, not remove one	§9.3
25 Oct	Report variant v3: "✓" only when the wording on the legal-state date is confirmed; every tool call of a judge logged	§9.3
25 Oct	Scope field describes the place of the error, not its effect on the score (B.6)	field filled in 49 pilot cards

The rules were changed after reading the answers, including three changes to A6 between 9 and 11 October 2026. For this reason the set is a calibration set: it illustrates the criteria and is not a ranking of models. The specification version is recorded in every card.

Appendix C. Question set

The ten questions, their answer keys and the reference cards will be published with the methodology once the cards have passed review (Statements); the text of the questions is therefore not repeated here. The table gives the domain, type, trap, expected calibration, key version and topic of each question in the W0/W1 set. Questions were asked in Polish, as typed by a lawyer, with spelling errors and abbreviations left in place. Every key has a legal-state date of 1 October 2026 and contains: the core answer, required legal bases, reference wording of the provisions copied from the official text, rubric elements, penalties, expected calibration and a field recording what could not be verified. A judgment enters the key only with its thesis copied from the text of the judgment and the database and scope of reading recorded; commentaries are not a source of a thesis.

Question	Domain	Type	Trap	Expected calibration	Key version	Topic
P01	petty offences	computation	yes	categorical	v3 of 2026-10-18	statutory threshold between a petty offence and a crime against property; trap: the former link to the minimum wage
P04	administrative procedure	closed	yes	categorical	v3 of 2026-10-18	permit for the acquisition of real estate by a foreign national: finality versus legal validity of the decision; trap: "ruling" and "clause" in the question
P05	civil procedure	closed	yes	categorical	v2 of 2026-10-10	service of court documents on a natural person through the court information portal; trap: outdated premise
P06	civil procedure	closed	no	conditional	v3 of 2026-10-10	rejection of an interlocutory appeal by a court of appeal and the available remedy (conditional on the type of appeal)
P07	civil procedure	closed	no	categorical	v2 of 2026-10-10	"principal reasons" of a ruling versus written justification; admissibility of the motion for reasons
P08	AI Act (EU)	verification	yes	conditional	v4 of 2026-10-18	definition of a deepfake under the AI Act, put to the system as a claim to verify; trap: a narrowed definition
P09	copyright	closed	no	conditional	v4 of 2026-10-18	transfer of copyright: fields of exploitation and written form (conditional)
P10	copyright	closed	yes	conditional	v3 of 2026-10-18	the same issue as P09, with a trap in the premise
P12	civil procedure	closed	yes	conditional	v1 of 2026-10-10	enforcement clause against the debtor's spouse; trap: "without limitation"
P13	civil procedure	closed	no	categorical	v1 of 2026-10-10	order of payment on a bill of exchange: stay of enforcement versus the order as a title of security

Five questions (P05, P06, P07, P12, P13) formed the pilot described in §6; five (P01, P04, P08, P09, P10) were added on 18 October 2026 to cover other fields. Keys P09 and P10 were revised after the run (18 October: conditional core, Supreme Court line as a rubric element); 20 cards were re-scored under the new keys before the judges ran on them.

Appendix D. Model judges: procedure and results

D.1 Inputs and outputs of a judge card

Element	Content
Input 1	scoring rules: blocks A and B of the specification, boundary examples, recorded decisions
Input 2	the answer key of the question in full: core, required bases, reference wording, rubric, penalties, expected calibration, legal-state date
Input 3	the text of the answer, without the name of the system
Instruction	check every cited provision and judgment in the official sources by web search (ISAP and ELI for Polish acts, EUR-Lex for EU law, SAOS or the Supreme Court website for judgments); compare the unit number and its content literally; do not rely on memory; list every citation that could not be checked
Output	a structured card: score per criterion, class of each error (F, T, W, S), binary label "hallucinated", list of checked citations with source and verdict, short justification
Logged	number of search queries; the queries (Vertex AI from the start, Codex from 25 October 2026); tokens; time
Measure	Cohen's κ between the judge and the reference card on the label "hallucinated"; false alarms (judge: yes, card: no) and misses (card: yes, judge: no)

D.2 Agreement, searches and declared unchecked citations

Judge	Set	n	Agreement	κ	False alarms	Misses	Searches per card	Cards without search	"Unchecked" per card
Gemini 4 Argon	extension (P01, P04, P08–P10)	79	95 %	0.90	1	3	12.2	0	0.01
Gemini 4 Argon	pilot (P05–P07, P12, P13)	68	99 %	0.97	0	1	16.2	1	0.01
Sol 6.1	extension (P01, P04, P08–P10)	80	90 %	0.80	7	1	11.2	0	0.36
Sol 6.1	pilot (P05–P07, P12, P13)	80	92 %	0.84	6	0	13.5	0	0.38
Gemini 3.1 Pro	extension (P01, P04, P08–P10)	80	90 %	0.80	0	8	4.3	25	0.00
Gemini 3.1 Pro	pilot (P05–P07, P12, P13)	79	86 %	0.71	5	6	7.6	9	0.08
Gemini 3.8 Flash	extension (P01, P04, P08–P10)	80	79 %	0.58	1	16	4.7	9	0.01
Gemini 3.8 Flash	pilot (P05–P07, P12, P13)	80	91 %	0.82	2	5	6.8	4	0.00
Astra 6	extension (P01, P04, P08–P10)	6	100 %	1.00	0	0	10.3	0	0.17
Astra 6	pilot (P05–P07, P12, P13)	43	98 %	0.95	1	0	10.2	0	0.37

Figures 6 and 7 in §9.1 show these figures with their 95 % intervals and against the number of searches.

D.3 Misses by provider of the evaluated model

Judge	Anthropic answers: missed / recorded	Google answers: missed / recorded	OpenAI answers: missed / recorded
Gemini 4 Argon	1 / 44 (2 %)	2 / 24 (8 %)	1 / 12 (8 %)
Sol 6.1	0 / 49 (0 %)	1 / 29 (3 %)	0 / 13 (0 %)

Judge	Anthropic answers: missed / recorded	Google answers: missed / recorded	OpenAI answers: missed / recorded
Gemini 3.1 Pro	6 / 48 (12 %)	7 / 29 (24 %)	1 / 13 (8 %)
Gemini 3.8 Flash	6 / 49 (12 %)	13 / 29 (45 %)	2 / 13 (15 %)

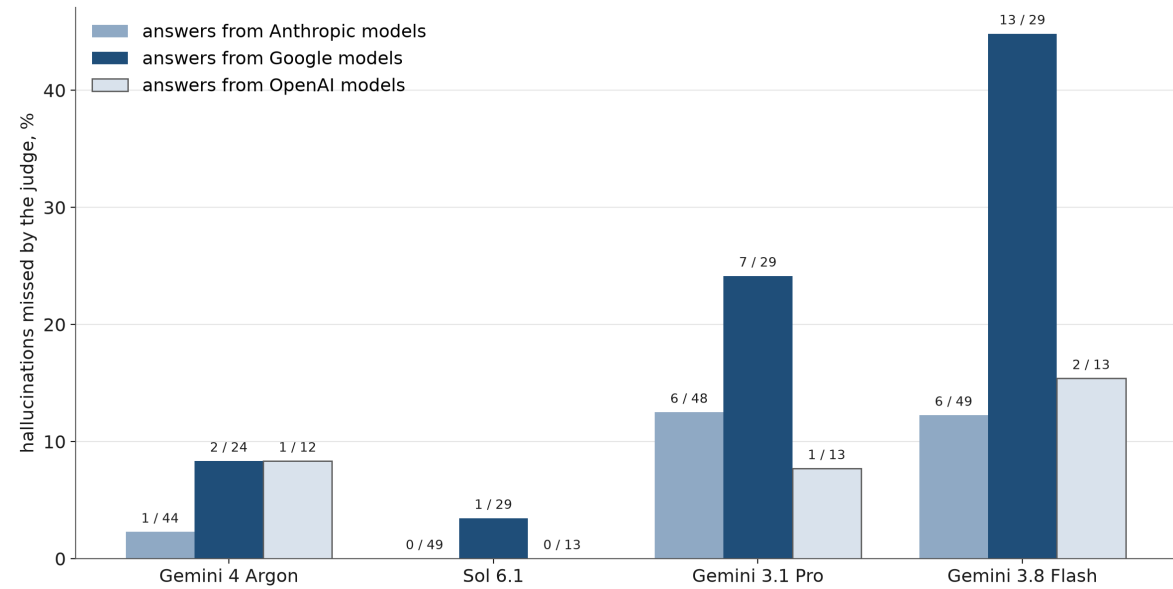


Figure A1. Hallucinations recorded in the reference cards and missed by the judge, by provider of the evaluated model; all ten questions. Google answers in this set contain many class *W* errors, which all judges detect less well, so the difference may reflect the type of error rather than the model family.

D.4 Panels of judges

On the 79 extension cards scored by all four judges, Gemini 4 Argon alone reaches κ 0.90. The table gives the rules we tested. The rules were chosen on the same data, so the gain of 0.02 is a fit, not a result.

Rule	κ	False alarms	Misses
Gemini 4 Argon alone	0.90	1	3
hallucination if Gemini 4 Argon or Gemini 3.1 Pro reports it	0.92	1	2
majority of three (Gemini 4 Argon, Sol 6.1, Gemini 3.1 Pro)	0.90	—	—
any of the four judges	0.82	7	0

D.5 Confusion matrices and uncertainty for every κ reported in the paper

Every κ in §6 and §9 is a summary of a 2×2 table: reference card and judge both "hallucinated" (TP), judge only (FP, a false alarm), card only (FN, a miss), neither (TN). The table below gives these four counts for every comparison, with the number of positives in the reference cards, raw agreement, κ , and a 95 % interval. The interval is a percentile bootstrap over questions: whole questions with all their cards are drawn with replacement (4,000 draws), which respects the dependence between the 13 or 16 answers to one question. With five to eight questions per set the intervals are wide, and this is the honest width. Where two intervals overlap, the text treats the difference as unresolved. The bootstrap is itself coarse at this size: with five questions there are few distinct resamples, and an interval of 1.00–1.00 means only that no question contained a disagreement, not that the agreement is certain. The number of positives in the cards is given because κ depends on it: the same raw agreement gives a lower κ when positives are rare. All rows use the same version of the reference cards, closed on 25 October 2026; the pilot figures in §6.2 (κ 0.81–0.93) refer to the cards before the author's review and are reported separately for that reason. The script that produces the table reads the cards and the judge files directly.

Judge / rule	Set	n	Positives in cards	TP	FP	FN	TN	Agreement	κ	95 % interval (bootstrap by question)
Astra 6	pilot (§6.2, §9.1)	65	43	43	0	0	22	100 %	1.00	1.00–1.00
Sol 6.1	pilot (§6.2, §9.1)	65	43	43	0	0	22	100 %	1.00	1.00–1.00
Gemini 3.1 Pro	pilot (§6.2, §9.1)	65	43	41	0	2	22	97 %	0.93	0.87–1.00
Gemini 3.8 Flash	pilot (§6.2, §9.1)	65	43	39	1	4	21	92 %	0.83	0.75–0.94
Gemini 4 Argon	pilot (§6.2, §9.1)	62	40	39	0	1	22	98 %	0.97	0.89–1.00
Gemini 4 Argon	extension (§9.1)	79	41	38	1	3	37	95 %	0.90	0.80–0.97
Sol 6.1	extension (§9.1)	80	42	41	7	1	31	90 %	0.80	0.59–0.95
Gemini 3.1 Pro	extension (§9.1)	80	42	34	0	8	38	90 %	0.80	0.75–0.85
Gemini 3.8 Flash	extension (§9.1)	80	42	26	1	16	37	79 %	0.58	0.40–0.70
Gemini 4 Argon alone	extension, cards with all four judges (§9.1)	79	41	38	1	3	37	95 %	0.90	0.80–0.97
Argon or Gemini 3.1 Pro	extension, cards with all four judges (§9.1)	79	41	39	1	2	37	96 %	0.92	0.87–0.97
majority of Argon, Sol 6.1, Gemini 3.1 Pro	extension, cards with all four judges (§9.1)	79	41	38	1	3	37	95 %	0.90	0.86–0.95
any of the four	extension, cards with all four judges (§9.1)	79	41	41	7	0	31	91 %	0.82	0.61–0.97
Gemini 4 Argon, without report	cards covered by v1 binding (§9.3)	54	35	34	0	1	19	98 %	0.96	0.88–1.00
Gemini 4 Argon, with report v1 binding	cards covered by v1 binding (§9.3)	54	35	31	0	4	19	93 %	0.85	0.76–0.95
Gemini 4 Argon, without report	cards covered by v2 auxiliary (§9.3)	115	62	59	0	3	53	97 %	0.95	0.90–0.98
Gemini 4 Argon, with report v2 auxiliary	cards covered by v2 auxiliary (§9.3)	115	62	56	1	6	52	94 %	0.88	0.86–0.91
Gemini 4 Argon, without report	cards covered by v3 status (§9.3)	116	63	60	0	3	53	97 %	0.95	0.90–0.98
Gemini 4 Argon, with report v3 status	cards covered by v3 status (§9.3)	116	63	58	2	5	51	94 %	0.88	0.86–0.91
Gemini 4 Argon, without report	cards shared by v2 and v3 (§9.3, §9.5)	115	62	59	0	3	53	97 %	0.95	0.90–0.98
Gemini 4 Argon, with report v2 auxiliary	cards shared by v2 and v3 (§9.3, §9.5)	115	62	56	1	6	52	94 %	0.88	0.86–0.91
Gemini 4 Argon, with report v3 status	cards shared by v2 and v3 (§9.3, §9.5)	115	62	57	2	5	51	94 %	0.88	0.85–0.91
Sol 6.1, without report	cards covered by v1 binding (§9.3)	61	41	41	3	0	17	95 %	0.88	0.81–0.96
Sol 6.1, with report v1 binding	cards covered by v1 binding (§9.3)	61	41	36	0	5	20	92 %	0.83	0.75–0.93
Sol 6.1, without report	cards covered by v2 auxiliary (§9.3)	119	68	67	8	1	43	92 %	0.84	0.70–0.96
Sol 6.1, with report v2 auxiliary	cards covered by v2 auxiliary (§9.3)	119	68	67	1	1	50	98 %	0.97	0.92–1.00

Judge / rule	Set	n	Positives in cards	TP	FP	FN	TN	Agreement	κ	95 % interval (bootstrap by question)
Sol 6.1, without report	cards covered by v3 status (§9.3)	128	73	72	11	1	44	91 %	0.80	0.68–0.93
Sol 6.1, with report v3 status	cards covered by v3 status (§9.3)	128	73	70	4	3	51	95 %	0.89	0.82–0.95
Sol 6.1, without report	cards shared by v2 and v3 (§9.3, §9.5)	119	68	67	8	1	43	92 %	0.84	0.70–0.96
Sol 6.1, with report v2 auxiliary	cards shared by v2 and v3 (§9.3, §9.5)	119	68	67	1	1	50	98 %	0.97	0.92–1.00
Sol 6.1, with report v3 status	cards shared by v2 and v3 (§9.3, §9.5)	119	68	65	2	3	49	96 %	0.91	0.85–0.98

Three readings follow. First, on the extension set the interval of Gemini 3.8 Flash (0.40–0.70) does not overlap with that of Gemini 4 Argon (0.80–0.97); the other differences between single judges are within the intervals. Second, the panel rules differ from Gemini 4 Argon alone by at most 0.02 and all their intervals overlap. Third, for Sol 6.1 the intervals with the auxiliary report (0.92–1.00) and with the status variant (0.85–0.98) overlap, as do the intervals without the report (0.70–0.96) and with the auxiliary report; the direction of the change is consistent across the comparisons, but its size is not resolved with eight questions.

Appendix E. The verification layer

E.1 Two steps before the judge

Step	Who	Input	Output
Citation check	code only, no model	citations extracted from the answer; for a provision: the wording of the unit on the legal-state date and earlier wordings (6, 12, 24, 60, 120 and 180 months before) from the official texts in ISAP/ELI; for a judgment: the case number searched in SAOS, then in a read-only case-law database; quotations compared literally	class F, W or S, or a verification status
Support check	a model with restricted input	one claim that the answer attributes to one source, and the text of that source only	"supports" (only with a verbatim quotation from the source, whose presence the code then verifies), or class T, or a status
Verification report	code	the two steps combined	for each citation a class (✓, F, T, W, S) or a status, with the official wording, its version and date

E.2 Verification statuses

A status is not an error of the model. A failed check does not show that the model was wrong, and the two outcomes are never given the same label.

Status	Meaning
source text unavailable	the official text could not be retrieved (for example, SAOS unavailable, consolidated text only as a scanned PDF)
wording on the legal-state date not established	the unit exists but its wording on the date could not be retrieved; earlier or later wording only
ambiguous case number	the same case number is used by different courts and the answer does not say which
act not established	the answer cites a unit without naming the act and the default act of the key does not apply
"in conjunction with" chain	a chain of provisions treated as one citation, not checked as separate claims
judgment not found	the case number is absent from both databases; absence of a record is not evidence of a fictitious judgment

E.3 Report variants given to the judge

Variant	Instruction to the judge	Tested on
v1, binding	base the hallucination verdict on the classes in the report	4 questions (P05, P06, P07, P13)
v2, auxiliary ("a report may add an error, it may not remove one")	a class F, T, W or S from the report is binding unless the judge quotes the official source against it; a "✓" confirms only that the unit exists and is not a verdict on content or date; the judge must still compare every citation with the source	8 test questions
v3	as v2, but a citation whose wording on the legal-state date was not established is shown with that status instead of "✓", and the judge is told to assess it itself	8 test questions

E.4 Results on the test questions

Measure	23 Oct (frozen prompts)	24 Oct (input fixes)	24 Oct (two rules added to the prompt)
sensitivity, any class (147 errors in the cards)	84 (57 %)	87 (59 %)	94 (64 %)
sensitivity, same class	n/a	66 (45 %)	74 (50 %)

Measure	23 Oct (frozen prompts)	24 Oct (input fixes)	24 Oct (two rules added to the prompt)
T, any class (83)	39 (47 %)	49 (59 %)	50 (60 %)
W, same / any class (44)	n/a / 25 (57 %)	17 / 28 (39 / 64 %)	24 / 32 (55 / 73 %)
S, same / any class (14)	n/a / 11	0 / 4	2 / 6
F, same / any class (6)	n/a / 5	0 / 6	0 / 6
false alarms on 665 citations accepted by the cards	8.0 % support check + 4.1 % citation check	1.8 %	1.8 %
cost of the run (Vertex AI)	USD 8.09	USD 7.87	USD 7.09

Sources: the three run reports of the verification layer (23 and 24 October 2026) in the benchmark repository. The 24 October runs reuse the citation check and cached support verdicts where the input did not change.

Figure 8 in §9.2 shows the three runs. The fall in false alarms came from four fixes to input processing (superscript letters in unit numbers, default act, same case number in different courts, "in conjunction with" chains), not from the prompts.

E.5 Judge with and without the verification report

Judge	Variant	Questions	n (cards shared with the judge without report)	κ without report	κ with report	False alarms without → with	Misses without → with
Gemini 4 Argon	v1 binding	P05, P06, P07, P13	54	0.96	0.85	0 → 0	1 → 4
Gemini 4 Argon	v2 auxiliary	P01, P04, P05, P06, P07, P08, P12, P13	115	0.95	0.88	0 → 1	3 → 6
Gemini 4 Argon	v3	P01, P04, P05, P06, P07, P08, P12, P13	116	0.95	0.88	0 → 2	3 → 5
Sol 6.1	v1 binding	P05, P06, P07, P13	61	0.88	0.83	3 → 0	0 → 5
Sol 6.1	v2 auxiliary	P01, P04, P05, P06, P07, P08, P12, P13	119	0.84	0.97	8 → 1	1 → 1
Sol 6.1	v3	P01, P04, P05, P06, P07, P08, P12, P13	128	0.80	0.89	11 → 4	1 → 3

Figure 9 in §9.3 shows the same comparisons. Variant v3, which replaces "✓" with a status when the wording on the date was not established, left Gemini 4 Argon unchanged and lowered Sol 6.1 relative to v2; the difference rests on two cards and is reported as a working hypothesis.

Appendix F. Error classes and origin in the W0/W1 set

The 160 answers of the W0/W1 set (Appendix A.3) are scored under the current specification. Of them 91 carry the label "hallucinated" (51 of 80 in W0, 40 of 80 in W1) and 46 are correctly grounded (19 in W0, 27 in W1). 57 of the 91 hallucinated answers have an error in the core and 34 only in a peripheral remark. The table counts answers containing each class; one answer may contain several.

Class	W0 (80 answers)	W1 (80 answers)	Total
T — content not in the source	42	24	66
W — wording from another date	18	17	35
F — source or unit does not exist	4	4	8
S — distorted identifier	4	5	9
answers with at least one hallucination	51	40	91

Figure 4 in §6.2 shows the classes by condition.

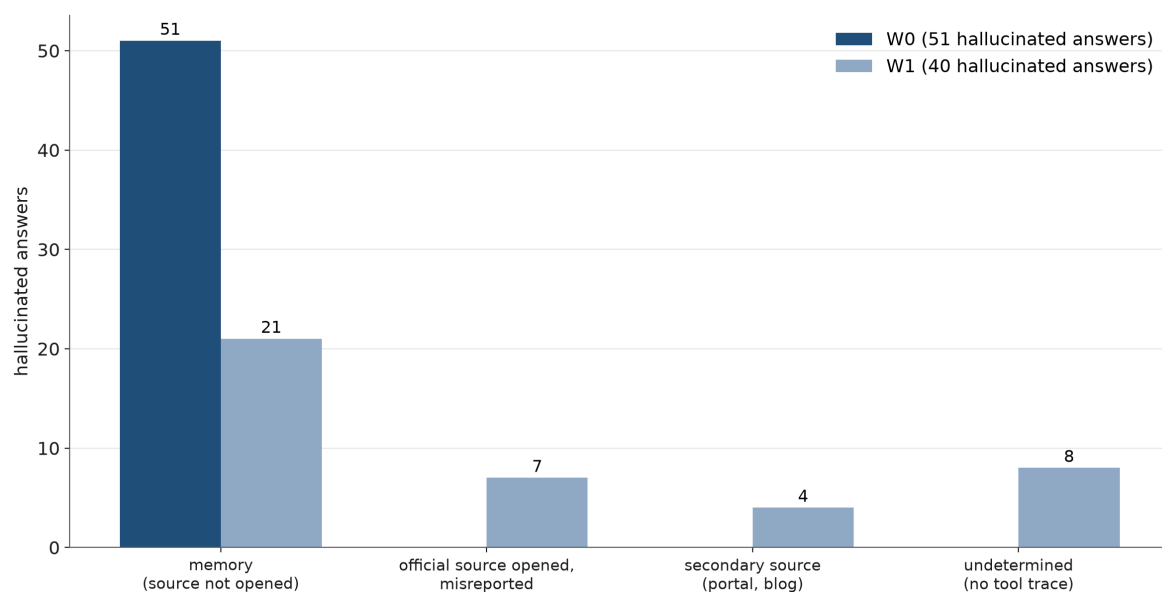


Figure A3. Origin of the hallucinated answers (Appendix B.5). In W0 every error comes from memory by construction. In W1, 21 of 40 hallucinated answers were written without opening the source, 7 after opening the right official source, 4 from a secondary source, and 8 cannot be determined because Claude Code keeps no tool trace.

Across the 160 answers the cards checked 1436 of 1454 citations against the official source; 1 answer cited nothing. The verification coverage of each card is recorded so that silence can be told apart from accuracy.

Appendix G. Cost

G.1 Price basis

Costs are API-equivalent amounts at the public price lists of 18 October 2026 (OpenAI, Anthropic, Google Vertex AI), standard tier, no batch discounts. For Codex and Claude Code runs through a subscription the amount is an equivalent, not an invoice; for Vertex AI it is the actual cost at the promotional prices then in force (Gemini 4 Argon and Gemini 3.8 Flash at 50 % of list price until 30 November and 31 December 2026 respectively). Reasoning tokens are priced as output tokens. Web searches by Gemini models (Grounding with Google Search) are priced at the list price of USD 0.014 per query; the first 5,000 queries a month are free of charge, so the invoice was lower than the sum below. Searches by Codex and Claude are priced at USD 0.01 per call where the count is recorded.

G.2 Full cost of the study

The table lists every model expense in this study: the evaluated answers, the model judges on the cards of specification 1.0.1, and the verification layer with all development and test runs. Judge and report costs are computed from the tokens recorded in the judge files. The amounts are a lower bound: judges in specification 1.0.1 did not record reasoning tokens, W1 runs of Claude and Gemini models have no complete tool record, and Codex runs through a subscription recorded no tokens. The table does not include the lawyer's time or the Claude model that filled in the cards.

Part	Item	Cards	Tokens USD	Searches	Searches USD	Total USD
Answers	160 answers of eight models (W0 and W1)	160	≥ 15.78	—	0.00	≥ 15.78
Model judges	Gemini 4 Argon, specification 1.0.1	147	≥ 6.18	2072	29.01	≥ 35.19
Model judges	Gemini 3.1 Pro, specification 1.0.1	159	≥ 5.73	945	13.23	≥ 18.96
Model judges	Gemini 3.8 Flash, specification 1.0.1	160	≥ 1.91	913	12.78	≥ 14.69
Model judges	Sol 6.1, specification 1.0.1	160	n/a	n/a	n/a	n/a (Codex subscription; tokens not recorded)
Model judges	Astra 6, specification 1.0.1	49	n/a	n/a	n/a	n/a (Codex subscription; tokens not recorded)
Verification layer	citation check (code, public sources)	—	0.00	—	0.00	0.00
Verification layer	support check (6 runs: tests, development on P09–P10, test on 8 questions, fixes)	—	26.38	—	0.00	26.38
Verification layer	judge card with report, Gemini 4 Argon v1	64	6.57	193	2.70	9.28
Verification layer	judge card with report, Gemini 4 Argon v2	125	15.97	571	7.99	23.96
Verification layer	judge card with report, Gemini 4 Argon v3	127	17.07	652	9.13	26.20
Verification layer	judge card with report, Sol 6.1 v1	61	n/a	n/a	n/a	n/a (Codex subscription)
Verification layer	judge card with report, Sol 6.1 v2	119	n/a	n/a	n/a	n/a (Codex subscription)
Verification layer	judge card with report, Sol 6.1 v3	128	n/a	n/a	n/a	n/a (Codex subscription)
Verification layer	smoke tests	—	0.26	—	0.00	0.26
Total			≥ 95.85		74.84	≥ 170.70

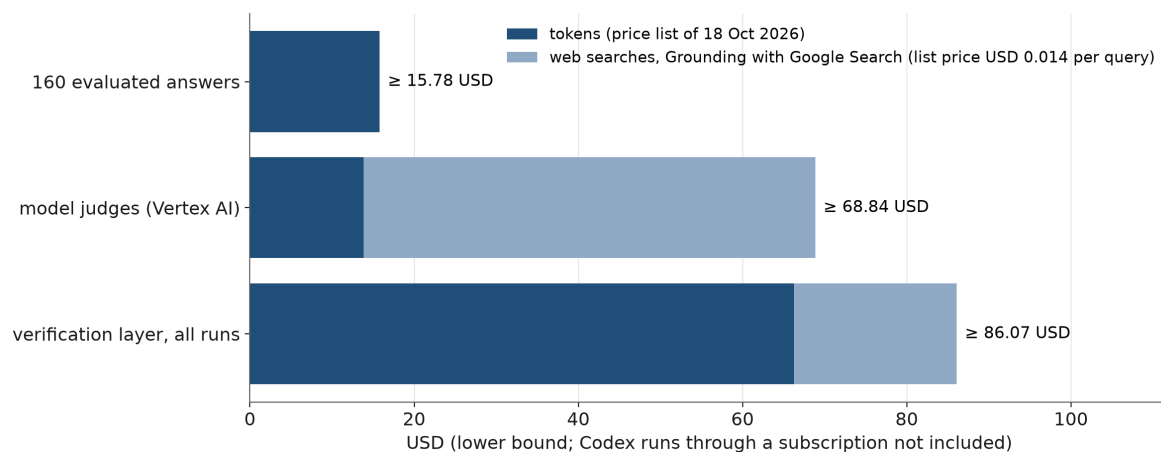


Figure A2. Full cost of the study by part (USD, lower bound). Evaluation cost several times more than generating the evaluated answers; for the Gemini judges the searches, at list price, cost more than the tokens.

The cost of one correctly grounded answer in the W0/W1 set ranges from USD 0.042 (Sol 6.1, W0) to at least USD 2.085 (Gemini 4 Argon, W1); for models with no correctly grounded answer the metric is undefined. A run with web search cost 1.4 to 7.7 times the run without tools for the same model, with a change in the number of correctly grounded answers between -1 and +4.

Statements

Data availability. The ten questions used in this paper, their answer keys, the scoring specification and the reference cards will be published together with the methodology once the cards have passed review. The remaining questions of the pool are withheld to keep them out of future training data and will be released with later versions of the benchmark. Until then the material is available to researchers on request. The reference cards are versioned by the date on which the set was closed; every figure in this paper refers to the version of 25 October 2026 unless the text says otherwise, and later versions will be published with a change log.

Competing interests. The author is a partner at KTZR, a law firm which develops its own legal research assistant. That system took part in the pilot run and is excluded from all numbers in this paper. Since 17 October 2026 systems developed by KTZR are not evaluated in this benchmark at all, because the evaluator would also be the author of the evaluated system. OKF-Legal, discussed in §10, is the author's own project, published under the Apache 2.0 licence; the comparison in §10 is therefore not an independent assessment of it. The study received no external funding and was financed by the author; all systems were accessed through standard paid subscriptions or cloud accounts. The author has no financial relationship with the providers of the evaluated systems.

Use of AI. Language models were used as scorers and judges, as described in §6.1 and §9. Every scoring card was first filled in by a Claude model and then reviewed and closed by the author. LLMs were used for language editing; the author is responsible for the content.

How to cite

Piotrowski A. (2026). *"Hallucination" Is Not Enough: Why Legal LLM Benchmarks Should Measure "Correctly Grounded"*. LexAlpha Lab working paper v1.2, 26 October 2026. <https://doi.org/10.5281/zenodo.23438195>