

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

Adam Piotrowski · LexAlpha Lab, KTZR · 17 October 2026 · working paper, v1.1 · DOI: [10.5281/zenodo.23438196](https://doi.org/10.5281/zenodo.23438196)

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 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". The most influential 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 transfer one-to-one to civil-law systems. 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 a wrong statutory content 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 published benchmark with numbers. I propose to replace the binary question "did the model hallucinate?" with "is the answer correctly grounded?", defined as: the cited designation exists, the content attributed to it is really there, and the wording was in force on the 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.

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

1. Introduction

At the beginning of our work on a legal benchmark I was impressed by the Stanford paper "Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools" [1]. Its definition of hallucination seemed precise and, as I then thought, universal.

After a few months of building our own benchmark for LLMs in Polish law I am not so sure anymore. The definition is not wrong, but in civil law (and in Polish law in particular) it does not apply to every case, and in practice it mixes two different things:

- the model's correctness in recalling the proper legal basis of its statement,
- the groundedness of the statement itself.

For Poland the starting point is simple. As far as I could establish, there is no public benchmark of legal hallucination in Polish law built on a generally accepted scientific method. There are two 2026 rulings of the Supreme Administrative Court (NSA) naming "AI hallucinations" in lawyers' briefs [8], and one study on the public procurement (KIO) exam in which a GPT-4o judge overrated the answers by 51 to 77 points against the human examination board [9] – a useful reminder of why the 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 true facts of the world. The paper focuses on the third. Factual hallucination is then decomposed into 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 what follows: a correct answer with no citation at all is not a hallucination in that framework, only "incomplete".

The authors also stress that "grounded" in their legal sense means grounded in the actual governing law, not in whatever document the system happened to retrieve, which is the computer-science sense of the word.

3. Three ways models ground their answers

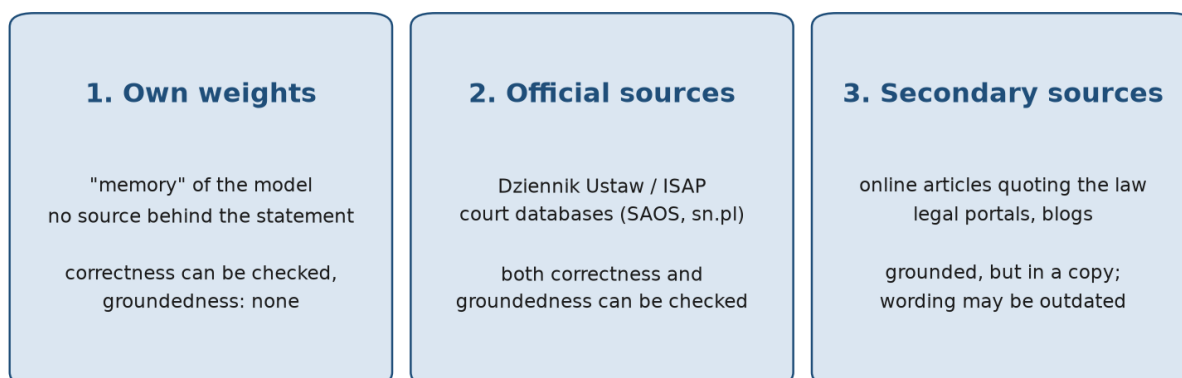
Groundedness is, in my view, the part that needs more attention. At this stage of our evaluations we observe three practices:

1. grounding on the model's own weights and "memory";
2. grounding on official, state sources of law (in Poland: Dziennik Ustaw, ISAP, court databases such as SAOS);
3. grounding on unofficial sources, such as online articles in which the law is quoted.

The third practice is "grounded" only in the computer-science sense. Our 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 bucket in the paper, four problems for a lawyer

There are "edge cases" which in the Stanford framework mostly land in one bucket ("misgrounded"), but which for a practitioner are very different things:

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

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; the paper's "misgrounded" covers the other three.

In such cases groundedness overlaps with correctness. Take the paper's own example: "the right to same-sex marriage is protected, *Miranda v. Arizona*" is coded as correct but misgrounded. Now take the opposite case, a correct article number with content which is not in the provision: is it incorrect (a false statement about the law) or misgrounded (the cited source does not support the claim)? Both at once. The paper resolves it by order: groundedness is assessed only for correct responses, so correctness decides first. In the final label it does not matter anyway, because incorrect or misgrounded equals hallucination. My conclusion is that these two rubrics need to be scored and reported separately. In our cards the substance of the answer and the grounding of every citation are two different scores.

5. Common law vs civil law: a wrong citation does not weigh the same

There is one more reason why the framework does not transfer one-to-one to Poland. It was built for US law, where case law is the source of law and a wrongly cited judgment is the core 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: what matters first is the content of the provision, and the judgment is usually an interpretation of it. So in our cards a wrong case number with a correct thesis is a smaller problem than a correct article number with content which is not there. Both are "misgrounded" in Stanford's terms, but for a Polish lawyer the second one is the dangerous one, and it is also the most frequent one in our results.

Two civil-law studies from 2026 report the same pattern from the other side. On the Japanese bar exam, 67% of the precedents cited by models were hallucinated against 15% of the statutes [5]. On a Brazilian legal platform, 53% of case-law references could not be resolved against 18% of references to legislation [6].

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 date of the legal state.

All answers were collected on 10 October 2026; the date of the legal state 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 allows one. "Bare" models (10 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; it is 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).

A6 is binary. An answer gets 0 and the label hallucinated when at least one statement about a source of law is inconsistent with the official text on the date of the legal state. The type 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 where the judges disagree with the card.

6.2 Results

Under the binary rule 41 of 50 bare answers were labelled hallucinated, and 7 of 10 models failed on all 5 questions. The three systems with web search failed 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 bare answers the most common problem was not a made-up source but a correct designation with content which is not in the provision (type T), found in 35 answers. Old wording (W) came second with 13 answers, a fictitious source (F) appeared in 4 and a real source with a wrong identifier (S) in 1.

On the grounding side, of 15 answers from systems which could search, 12 reached an official source and 3 stayed only on secondary sources (legal portals, blogs). Adding the 50 bare answers, in 53 of 65 answers in the whole run the model worked only on its own weights or on secondary sources.

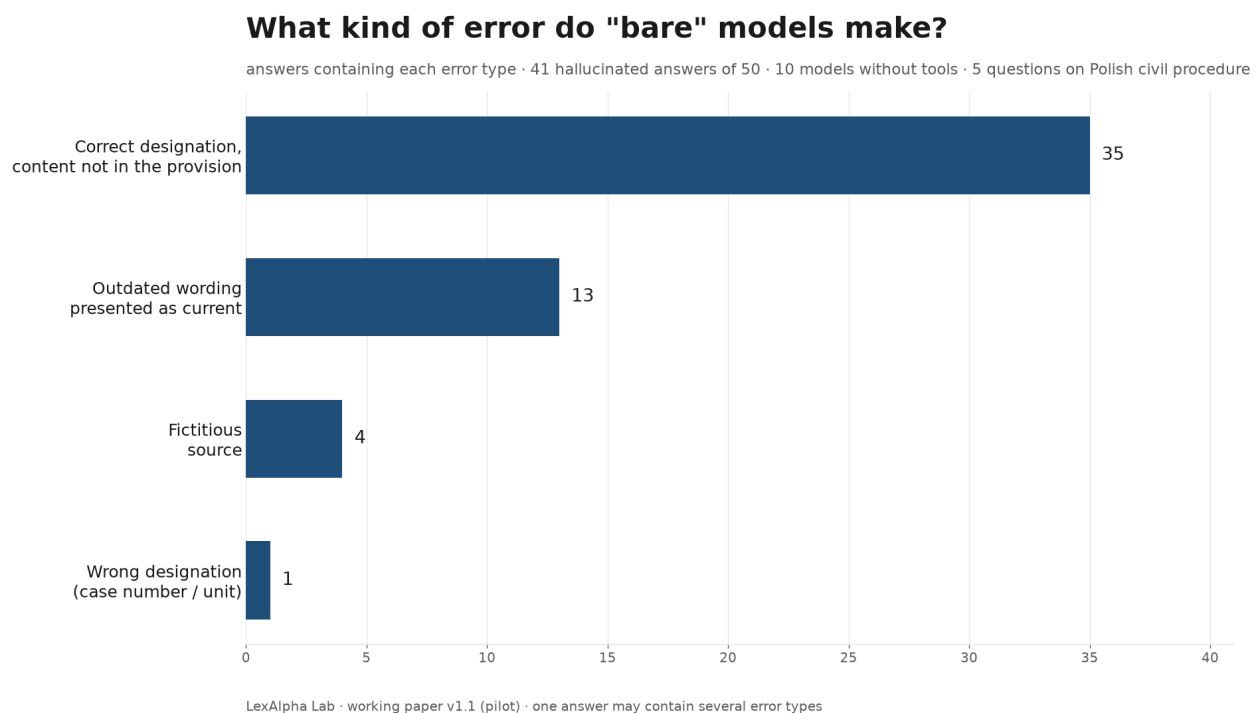


Figure 3. Error types in 41 hallucinated answers from 10 models without tools: number of answers containing each type (one answer may contain several).

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 comparison is not like for like: the first scorer of every card is itself a model, so this is agreement between models, checked by a human only where they disagree.

6.3 Limitations

This is a pilot. 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 cases where the judges disagreed with the card; κ is therefore reported before the review, because after it the agreement is partly the result of the author accepting the judges' view (it rises to 0.81–1.00). 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"

So perhaps the question is not "did the model hallucinate?" but "is the answer correctly grounded?". Stanford uses exactly this phrase, grounded "in the legal sense", in the actual governing law, and I think this is the direction for legal benchmarks. In our cards every cited provision is checked against the official text, and the answer is correctly grounded only if the designation is right, the content is really there and the wording is current. A single wrong unit (art. 1 instead of 1¹) breaks it, even if the gist was right. Checking that the wording was in force on the date of the question is, as far as I found, something no published benchmark with numbers does yet; ISAP gives that date for free.

A correct answer without any source is not a hallucination, but it is not correctly grounded either.

This is harsh, but the metrics for human lawyers do not transfer: a lawyer who writes § 1¹ instead of § 1² made a typo and knows what they meant; a model has no such memory of intent.

The Stanford paper offers one more reason to keep the "no source" label separate. The authors observed Westlaw stating propositions from overruled cases without citing them, probably suppressing "red-flagged" citations in a second pass. A statement without a source is exactly the one which is hardest to verify and most likely to be false.

8. Why the criteria have to be simple

The Stanford study was done 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 solid, but it describes Lexis and Westlaw as they were in spring 2024.

Meanwhile new models are released almost every month. If a benchmark needs months to evaluate, we end up reviewing models nobody uses anymore. To keep a benchmark current we need clear criteria which can be applied fast and repeated 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, but because it can be applied to a model released yesterday and defended. Under the binary verdict the cards keep per-citation labels (which provision, what type of error), so finer rates can be computed without changing the method.

This direction is not ours alone. 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.

There is one point where we openly 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 do use model judges, as described in §6.1, and we

report their agreement with the cards. I do not claim this replaces human scoring; it is the price of keeping up with monthly releases. What we are trying to build is a system in which a human has to review only a small share of the cards, those where the judges disagree or the risk of error is highest, because otherwise every benchmark will keep crawling at the pace of the slowest reviewer. This concerns scoring a benchmark, not legal practice: there every citation still has to be checked, as the SRA and CCBE guidance say [10].

9. What legal AI is actually good for right now

If I had to say where the real value of "legal AI" is today, it would not be 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 good enough to understand what a judgment actually 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 look for the sources. Knowing which provisions to look for needs a proper framework, and our numbers say the same: from memory 41 of 50 answers failed, while the three systems with web search failed in 2 of 15.

10. Conclusion

The edge cases listed here are only the ones we can distinguish at this moment. 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 a wrong statutory content in civil-law systems. I would welcome comparison with anyone measuring the same thing in other jurisdictions.

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 NHR).
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., LexAlpha Lab benchmark report (Polish; methodology of the scoring cards, criteria A/B, typology F/T/W/S), KTZR 2026, unpublished internal report, available from the author on request.
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, 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, 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.

Appendix A. Systems in the pilot run (r3)

All answers collected on 10 October 2026; date of the legal state: 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

Model judges (scoring specification 1.0.1): Sol 6.1 and Astra 6 (Codex CLI, with web search), Gemini 3.1 Pro and Gemini 3.8 Flash (Vertex AI, Google Search, temperature 0).

Statements

Data availability. At this moment the questions, answer keys and scoring cards are not released publicly, to keep the questions out of future training data; part of this material will be published with later versions of the benchmark. Until then they are available to researchers on request.

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. 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. 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 v1.1, 17 October 2026. <https://doi.org/10.5281/zenodo.23438196>