Position: Legal LLM Hallucination Should Be Evaluated as Failure of Legal Warrant

Maksym Taranukhin [1 2] **Vered Shwartz** [1 2 3]

Abstract

In this position paper, we argue that legal LLMs' hallucinations should be evaluated as a failure of legal warrant rather than as factual inaccuracy or citation failure. We define claim-authority warrant as the context-sensitive relation between a consequential legal claim and authority that exists, applies to the relevant jurisdiction, is current for the date of analysis, has the legal status represented by the system, and supports the proposition asserted. Warranted legal generation is the broader system behavior that answers, narrows, asks, warns, corrects a false premise, or abstains according to that relation. The falsifiable prediction is that warrant metrics reveal material failures that answer accuracy, citation existence, generic attribution, LegalHalBench-style statute relevance, and CitaLaw-style sentence-citation alignment can miss. We sharpen this claim with a side-by-side comparison item and a small, reproducible pilot over public-rule tests. We then specify benchmark records, claim boundaries, support labels, mixed response-policy scoring, risk weights, annotation reliability reporting, and jurisdiction-specific authority ontologies. The result is a concrete research agenda for evaluating legal AI systems by whether their consequential claims are licensed by law.

1. Introduction

Legal hallucination is often introduced as the invention of cases. That is the simplest failure, not the whole problem. In *Mata v. Avianca*, lawyers filed non-existent cases and quotations generated by ChatGPT, leading to Rule 11 sanctions (United States District Court for the Southern District of New York, 2023). More recent public incidents include filings or proposed orders with inaccurate citations, misstatements, fictitious or misattributed authority, and AI-assisted drafting failures (Freifeld & Scarcella, 2026; Scarcella, 2026). These events mix fabrication with a subtler defect: a legal proposition can be attached to a real source and still not be licensed by that source.

The research evidence points in the same direction. General-purpose LLMs hallucinate on legal knowledge questions, vary across courts and time periods, accept false premises, and often lack calibrated awareness of their errors (Dahl et al., 2024). AI legal research tools that use retrieval reduce but do not eliminate incorrect or misgrounded answers (Magesh et al., 2025). Legal retrieval benchmarks show that finding candidate authority remains difficult when answers depend on facts, exceptions, rule hierarchy, and source treatment (Zheng et al., 2025).

Therefore, this position paper argues that **legal LLM hallucination should be evaluated as a failure of legal warrant.** A legal claim is not reliable because it sounds plausible or carries a real citation. A blog post, dissent, district court case, agency FAQ, and controlling statute do not warrant the same kind of claim. Legal reliability depends on whether authority licenses a proposition under the relevant jurisdiction, date, forum, procedural posture, source status, and support relation. This view builds on argumentation theory and legal reasoning about authority (Toulmin, 1958; Schauer, 2009), but it changes ML evaluation. The object to evaluate is the relation between each consequential legal claim and the authority offered, retrieved, omitted, or shown to be unavailable.

Figure 1 shows an example that makes the target concrete. A user asks whether they have 30 days to appeal a federal civil judgment against the Department of Veterans Affairs. An answer citing Federal Rule of Appellate Procedure 4(a)(1)(A) for a 30-day rule uses a real, topical source, but it is overbroad. Rule 4(a)(1)(B) gives 60 days when the United States or a federal agency is a party. By contrast, a private-party version of the prompt would make Rule 4(a)(1)(A) the more relevant starting point (Administrative Office of the U.S. Courts, 2025).

[1]University of British Columbia, Vancouver, BC, Canada [2]Vector Institute, Toronto, ON, Canada [3]CIFAR AI Chair. Correspondence to: Maksym Taranukhin <maksymt@cs.ubc.ca>, Vered Shwartz <vshwartz@cs.ubc.ca>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

Figure 1. Example of a legal warrant for a federal appeal deadline. (a) The recorded facts make Rule 4(a)(1)(B)'s 60-day period applicable while Rule 4(a)(1)(A)'s 30-day period governs only the private-party counterfactual. (b) Although the 30-day authority exists and is topically relevant, it does not apply to the recorded facts or support the answer.

This paper makes four contributions, which also define its structure. First, it introduces the **C**laim-**L**aw **A**uthority **W**arrant (CLAW) framework (Section 2), which defines legal warrant as an operational target and removes a circular definition by separating claim-authority warrant from response-policy adequacy. Second, it maps recent legal benchmarks such as LegalHalBench and CitaLaw onto the warrant target, showing that they measure necessary components of warrant but do not by themselves establish it (Sections 2 and 7) (Hu et al., 2025; Zhang et al., 2025). Third, it gives a proof-of-concept annotation pilot that shows how citation existence and topicality can pass while the warrant fails (Section 3). Fourth, it supplies a concrete roadmap for warrant benchmarks and warrant-aware systems, including a minimum viable warrant suite (Sections 4 to 6).

2. The CLAW Framework

Evaluating warrant requires a precise statement of what is being checked, for which unit of text, and in which legal context. The CLAW framework uses *claim-authority warrant* for a ternary relation $W(c, a, k)$ among a claim c, an authority or marked absence of authority a, and legal context k. The context records jurisdiction, forum, date of analysis, procedural posture, user type, source corpus, and source-status ontology. A claim-authority pair is warranted when the source exists, the source supports the proposition, the source applies in the recorded legal context, the source is current for the date of analysis, and the source has the status represented by the system. *Response-policy adequacy* is separate. It asks whether the system should answer, narrow, ask for missing facts, warn, correct a false premise, or abstain. *Warranted legal generation* is the composite system goal: make only warranted consequential claims and choose a response policy suited to missing, weak, or contested warrants.

The unit is a *consequential legal claim*. We propose a master boundary rule (see details in Section A). A generated statement is consequential if changing its truth value would plausibly change a user's legal action, risk assessment, deadline, remedy, argument, right, duty, burden, forum choice, or need to seek professional help. For public-user tools, the test is whether a reasonable lay user might act differently because of the statement. For lawyer-facing tools, the test is whether the statement would need authority in a memorandum, brief, advice letter, or research note. Background definitions are consequential only when they are used to support advice, triage, or a legal conclusion. Domain guides should instantiate this master rule with examples and counterexamples.

Table 1 separates failures often collapsed into one hallucination rate, in contrast to coarser legal-AI risk taxonomies (Buchicchio et al., 2024). Its eight rows are diagnostic categories for stress-test design and reporting, not eight independent labels assigned to every span. Annotators instead record a compact schema for each claim-authority-context tuple: source existence and status, jurisdiction-time-posture metadata, one support label, a response-policy vector, and a risk weight. The categories are derived from those fields, which keeps annotation modular and permits later consolidation when dimensions are redundant. A system can have perfect citation existence and still fail attribution, source status, time, jurisdiction, or posture. It can also avoid false claims by refusing everything while withholding warranted information. Warrant evaluation therefore needs both neg-

Table 1. Legal-warrant failure modes and reasons common evaluation scores may miss them; a single answer may exhibit multiple modes.

Failure mode	What fails	Why common scores miss it
Existence	A case, statute, rule, quotation, docket, or citation does not exist.	Citation checks catch this, but often treat it as the whole problem.
Attribution	A real source is cited for a proposition it does not support.	The answer appears grounded because the source is genuine and topical.
Authority status	Persuasive authority, dicta, a dissent, a trial ruling, or secondary material is represented as controlling law.	Generic support evaluation does not consider the precedential value or legal authority of the source.
Jurisdiction	The answer imports law from the wrong forum, agency, court hierarchy, or legal system.	The proposition may be true somewhere else.
Temporality	The rule is obsolete, amended, overruled, stayed, or not yet effective.	Static labels hide the date for which a rule is valid.
Procedural posture	The answer ignores stage, burden, standard of review, remedy, or filing posture.	The statement may be abstractly true but unusable.
False-premise compliance	The system accepts an incorrect legal assumption rather than correcting it.	The answer can be coherent relative to the prompt while failing the legal need.
Epistemic overreach	The system gives confident advice when authority is missing, conflicted, or facts are unknown.	The answer may be partly true, but lacks warranted scope.

ative labels for unsupported claims and positive labels for useful narrowing. Table 2 compares this record with adjacent legal benchmarks. Epistemic overreach from Table 1 is operationalized below under response policy, since confident advice without supporting authority is a policy failure.

We predict that warrant metrics will produce the largest gaps over adjacent metrics for attribution, authority status, procedural posture, and response policy. A source can be real, relevant, and sentence-aligned while supporting only a narrower proposition; existing metrics often do not test whether a source is binding, persuasive, secondary, dicta, dissent, or negatively treated; and relevance or citation alignment can reward an answer that should instead ask, narrow, warn, correct, or abstain. Procedural-posture gaps should be especially large for deadlines, remedies, burdens, standards of review, and litigation stage. We expect medium-to-high gaps for jurisdiction and time, which dataset scope can conceal even though deployed systems answer across borders and dates, and smaller gaps for existence and topicality, which current legal citation benchmarks already target well.

These predictions can be audited on existing outputs by adding metadata and support labels. Attribution audits should relabel cited sentences as direct, inferential, partial, contradiction, no address, out of scope, or unsettled; authority-status audits should add source-type and treatment metadata and compare the force claimed with the source's local legal force. Jurisdiction-and-time audits should swap jurisdiction, effective date, or forum while keeping the topical source visible, and posture audits should record whether the output conditions the rule on posture and whether the cited authority applies to that posture. Response-policy audits should label answer, narrow, ask, warn, abstain, and correct acts as a policy vector and apply vetoes to unsafe unsupported claims. Existing statute-existence, relevance, and citation-validity checks remain shared submetrics.

The running example in Figure 1 appears successful under adjacent scoring objects even though its consequential advice remains unsupported. Citation existence passes because FRAP 4(a)(1)(A) exists, but existence alone cannot determine whether another provision controls the facts. LegalHalBench-style relevance is high because the rule concerns civil appeal deadlines, but topical relevance does not establish that the rule supports this user's deadline. CitaLaw-style sentence-citation alignment passes or partially passes because the narrower statement "Rule 4(a)(1)(A) says 30 days" is entailed, but sentence-level alignment can still pass when the final advice is broader than the cited proposition.

The warrant record instead fails both support and response policy because it asks which authority controls for the recorded party type and date. Here, Rule 4(a)(1)(B) governs when a federal agency is a party, so a warranted response should narrow the claim, cite the 60-day provision, and avoid an unqualified "yes." This divergence motivates the pilot below, which tests whether the separation persists across a broader set of stress tests.

3. A Pilot Annotation

To make the empirical claim concrete, we constructed and labeled a reproducible test. The six prompts are manually designed stress tests. Each prompt targets one or more failure modes in Table 1: near-miss authority (a real and topical source that governs a slightly different situation than the user's), false-premise compliance, temporal treatment, procedural posture, jurisdictional underspecification, and source-status overreach. The jurisdictional items reflect evidence that hallucination rates vary across places and jurisdictions for place-based legal queries (Curran et al.,

Table 2. Coverage of warrant dimensions in LegalHalBench, CitaLaw, and the proposed warrant record; "partial" denotes coverage of some instances without a required field for every consequential claim.

Dimension	LegalHalBench	CitaLaw	Warrant record
Source existence	Strong for named statutes through non-hallucinated statute rate.	Present through retrieved law articles and cases.	Required for every cited or offered authority, including quotations and pinpoint text.
Topical relevance	Strong through statute relevance.	Strong through retrieval and citation attachment.	Necessary but not sufficient. A topical source may still fail to support.
Claim support	Scored as legal claim truthfulness.	Scored through sentence-citation entailment and syllogism components.	Labeled per consequential claim as direct, inferential, partial, contradiction, no address, out of scope, or unsettled.
Jurisdiction and time	Partial when encoded by dataset scope.	Partial when encoded by the corpus and question.	Explicit fields in each record, with old and new law preserved as temporal tests.
Authority status	Limited for statute-centered items.	Partial through the law article versus precedent distinction.	Jurisdiction-specific source ontology with binding force, institutional rank, treatment, and role in reasoning.
Procedural posture	Partial through scenario text.	Partial through the circumstances component.	Required context field when posture changes, remedy, burden, deadline, or standard.
Response policy	Helpfulness or style can reward usable answers.	Style and correctness are measured.	Multi-label behavior score for answer, narrow, ask, warn, abstain, and false-premise correction.

2025). One bankruptcy item is inspired by published RAG audit examples where legal research tools overstated jurisdictionality from topical bankruptcy material (Magesh et al., 2025).

For each prompt, we manually wrote three short candidate outputs rather than generating them with a model. The output patterns are a default-rule answer with a real topical source, a broad refusal or generic disclaimer, and a warranted-narrowing answer. A warranted-narrowing answer gives only the part licensed by the available authority, states the condition under which it applies, asks for missing legally operative facts when needed, and warns against treating an unverified condition as settled. This produces 18 outputs. The 42 claim-authority pairs come from extracting every consequential claim in those outputs and linking each claim to the source it offers or implicitly relies on. Section E gives the prompts, output templates, outputs, and labels. The pilot does not evaluate any deployed model; it shows how candidate answers would be scored. It tests whether the proposed labels are operational and whether they separate warrant from adjacent metrics; Figure 2 reports the pass rates by measure and scoring unit.

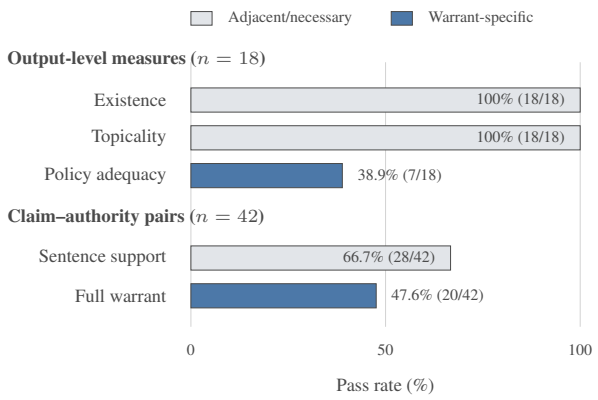

Figure 2. Pilot pass rates by scoring unit, with exact counts; the adversarial sample evaluates label separability rather than prevalence, and output-level and claim–authority-pair measures use different denominators.

Three observations follow. First, in this adversarial pilot, citation existence and topicality are deliberately easy to satisfy because every default answer cites a real, topic-adjacent source. The measured gap, therefore, isolates the harder dimensions: attribution, temporal or procedural fit, authority status, and response policy. Second, partial support is common. A source may support "ordinary civil appeals are due in 30 days" while not supporting "your appeal is due in 30 days." Third, broad refusal is not a solution. It avoids unsupported claims, but it fails to give warranted procedural information. The private-party near miss also shows that a default-rule answer can be adequate when the facts actually satisfy the default. This pilot is deliberately small, intended to make the falsifiable agenda testable. A future release can replace the hand-written outputs with outputs from real systems and ask whether the gap persists.

4. A Roadmap for Warrant Benchmarks

In this section, we turn the warrant target into a benchmark-building protocol and assess the feasibility and cost of that protocol. The goal is to evaluate generated legal answers by decomposing them into consequential claims, authorities, context metadata, support relations, response-policy acts, and user-risk weights. The same records can also support auxiliary modeling tasks, such as claim extraction or support-label prediction, but the primary benchmark use is open-system evaluation, i.e., a system produces an answer, annotators or validated tools build a warrant record for that answer, and scores are reported by dimension.

Grading pipeline. Evaluating an open system therefore follows an explicit workflow. The system under test produces an answer to the benchmark prompt. Consequential claims are then extracted from the answer, manually or with model assistance, and audited by a legally trained reviewer against the claim-boundary rule. Each audited claim is linked to the authority it offers or implicitly relies on, and the link is resolved against the frozen source snapshot for the item's analysis date. Finally, annotators assign the support and response-policy labels defined below. Model assistance can extend beyond extraction: an LLM-as-judge can propose support labels at low cost, but published audits of legal AI tools show that automated judges share the failure modes under evaluation, so judge proposals for high-risk claims require expert confirmation (Magesh et al., 2025, §5.3). The pipeline's expected errors are also predictable. False positives arise when a judge or annotator accepts topical but non-supporting authority or overlooks a defeating condition, and false negatives arise when a strict reading rejects legitimate inferential support or when the snapshot omits an authority the system validly relied on. Reporting audit rates for extraction and linking alongside label agreement makes these error sources visible.

A benchmark item should contain more than a prompt and an answer label. It should record the user scenario, jurisdiction, forum, date of analysis, procedural posture, user type, source corpus, gold issue, acceptable authorities, source status, support labels, response-policy labels, and risk weights. The scored object is a tuple (q, c, a, m, r), where q is the task context, c is a consequential claim, a is an authority or marked absence of authority, m is metadata about jurisdiction, date, forum, posture, and source status, and r is the response policy. The same surface answer can receive different labels when jurisdiction or date changes.

Support labels. Each consequential claim should be linked to an authority or to a label for which no legal authority is needed. Direct support means the source states the proposition, and its conditions are met. Inferential support means the source licenses the proposition through a stated legal inference, such as a definition, exception, or incorporated rule. Partial support means the source supports a narrower proposition or only one required condition. Contradiction means the source refutes the claim. No address means the source is topical but silent on the proposition. Out of scope means the source is legally inapplicable because of jurisdiction, time, status, or posture. Unsettled means reasonable authority conflicts or no controlling authority resolves the issue. Table 3 gives the decision rule and a typical legal example for each label. Existence itself admits strictness levels: a cited source may exist as a document while the pinpoint provision, quotation, or role it is cited for does not. Benchmark guides must state which strictness level their existence check uses; calibrating that choice is future work.

A good-faith legal argument can be warranted even when it is not legally correct in the sense of winning. The output must accurately state current law, mark the requested extension or change as an argument, and disclose contrary authority. The failure is not losing the argument. The failure is representing an analogy, dissent, policy preference, or proposed extension as controlling law.

Mixed response policies. The policy label should be a multi-label vector over answer, narrow, ask, warn, abstain, and correct, with definitions given in Table 9. A response can properly combine acts, such as answering a general process question, narrowing the deadline claim, asking for jurisdiction (Taranukhin et al., 2026), and warning that local rules may matter. Benchmark reports should compute micro and macro F_1 between the output's policy vector and the gold policy vector. They should then apply deterministic veto rules for unsafe behavior. For example, an output receives zero policy adequacy for a high-risk item if it states an unsupported deadline, eligibility rule, forum choice, waiver consequence, custody consequence, detention consequence, or immigration consequence as settled law. It receives capped credit when it asks a useful question but also overclaims, or when it refuses everything although a narrower warranted answer is available. These vetoes are pre-registered in the annotation guide, so they are part of the metric rather than post-hoc judgment.

Risk weights. User-risk weighted warrant error should use a pre-registered harm rubric rather than ad hoc weights. Low risk covers background statements unlikely to change the action. Medium risk covers claims that may affect planning or document preparation. High risk covers deadlines, forum choice, eligibility, waiver, custody, housing loss, detention, immigration status, criminal exposure, or irreversible procedural steps. Weights should be assigned by task designers and legally trained reviewers, reported by domain, and accompanied by sensitivity analysis. Validation against observed user harm is a later empirical task, and is not a prerequisite for useful benchmark construction.

Reliability. Annotation should be staged: record context,

Table 3. Decision rules and legal examples for support labels applied to claim–authority–context tuples.

Label	Decision rule	Typical legal example
Direct support	The source states the proposition, and all stated conditions are satisfied by the scenario.	A rule states the exact filing period for the recorded forum and party type.
Inferential support	The source licenses the proposition through an explicit legal inference using definitions, cross-references, or exceptions.	A statute defines a covered person and another section imposes the duty.
Partial support	The source supports a narrower proposition or only one element of the generated claim.	A default rule supports a deadline for ordinary civil cases, but not the federal-agency exception.
Contradiction	The source refutes the proposition or supplies a rule that makes it false in context.	Current law rejects a standard that the answer says controls.
No address	The source discusses the topic, but does not speak to the proposition.	An agency FAQ describes appeals generally but does not state the user's deadline.
Out of scope	The source is legally inapplicable because of jurisdiction, date, forum, posture, or source status.	A California trial order is offered as controlling law in a British Columbia housing matter.
Unsettled	Valid sources conflict or no controlling source resolves the claim.	Split authority on a novel statutory interpretation.

extract claims, link candidate authority, label support, then label response policy. Benchmarks should publish agreement by dimension rather than one aggregate number. Existence and pinpoint matching may have high agreement. Inferential support, source treatment, and posture may not. That pattern is informative because it tells developers which parts of the legal warrant need better metadata or clearer annotation guides. LegalBench and LexGLUE show that legal tasks can be collaboratively annotated and benchmarked, but warrant annotation should preserve disagreement for unsettled law rather than force false consensus (Guha et al., 2023; Chalkidis et al., 2022).

Warrant cards. Evaluation should produce a warrant card for each output, with the minimum fields listed in Table 4. A lawyer-facing interface may expose the full card. A public-facing interface may translate it into plain language. The card is not a new disclaimer. It is a compact representation of what the system actually knows, what it assumes, and which claims remain unsupported.

4.1. Feasibility, Cost, and Generalization

Scope. A first warrant benchmark should be narrow. Good early domains include federal appellate deadlines, state housing repairs, immigration form triage, benefit eligibility thresholds, and administrative appeal windows. These tasks are compact enough for expert review and consequential enough to reveal failures. A realistic first release might contain 250 prompts, three seed outputs per prompt, and 1,500 to 3,000 claim-authority pairs. This scale is diagnostic, not exhaustive of the long tail of legal phrasing, domains, jurisdictions, or user circumstances. Coverage should be reported by stress-test family and expanded with paraphrase, counterfactual near-miss, and cross-jurisdiction variants, while claim-extraction precision and recall are measured on fresh system outputs. Seed outputs provide reproducible reference material for comparing metrics, training auxiliary extractors, and stress-testing annotation. New systems are evaluated by applying the same extraction, linking, and support-label protocol to their own outputs.

Cost. A first-pass annotator can draft claim spans and source links in 10 to 20 minutes per prompt when the domain and source corpus are narrow. Legal review and adjudication may add 20 to 40 minutes for difficult items. A 250-prompt release therefore has a rough budget of 125 to 250 expert hours, plus setup time for source snapshots and annotation guides. These estimates come from a narrow pilot, and genuinely hard items may exceed them. This is comparable to recent legal benchmark construction efforts that already report substantial lawyer time, such as LegalHalBench's more than 200 hours of professional review (Hu et al., 2025). Costs can be reduced by model-assisted claim extraction, reusable authority graphs, and active sampling, but the final support and policy labels for high-risk claims should remain legally reviewed.

Generalization. The authority-status ontology must be jurisdiction-specific. The common-law labels in many U.S. examples, such as controlling, persuasive, dicta, dissenting, overruled, and superseded, are one instantiation. Civil-law systems may use code articles, regulations, constitutional decisions, administrative interpretations, jurisprudence constante, doctrinal commentary, and court decisions with different formal force. The general schema should therefore encode functional dimensions: source type, institutional rank, binding force, temporal effect, treatment status, and role in the reasoning. Local benchmark guides should map those dimensions to the legal system being tested. Cross-jurisdiction evaluation should compare whether systems respect the local ontology and not whether every system fits U.S. common-law categories.

Source snapshots. Legal materials change. A benchmark

Table 4. Minimum fields in a warrant card for a generated legal answer.

Field	Purpose
Legal context	Jurisdiction, date, forum, domain, user type, and posture.
Claim ledger	Consequential claims extracted from the answer.
Authority links	Sources offered for each claim, with pinpoint locations.
Support label	Direct, inferential, partial, contradicted, not addressed, out of scope, or unsettled.
Status metadata	Binding force, source type, treatment, and effective date.
Response policy	Answer, narrow, ask, warn, abstain, or correct.
Risk note	Why an unsupported claim could harm the user.

should store or identify the version of each source available at the analysis date. When the law changes, the old item can remain valid as a temporal test, and a new item can be added for the new date. Papers should state what metadata the system had access to. A system with proprietary citator treatment history should not be compared naively with one restricted to public text.

Reporting. Benchmark builders should resist a single leaderboard culture. Warrant data is most valuable when it reveals where systems break. A model might be strong on existence and weak on procedural posture. Another might ask good clarifying questions but omit contrary authority. Reporting these dimensions separately will make progress slower to summarize, but more useful for deployment. It will also reduce incentives to optimize for citation density or polished disclaimers at the expense of actual support.

5. A Roadmap for Warrant-Aware Systems

Benchmarks are only half of the agenda. This section turns the warrant target into design guidance for systems, with access to justice as the motivating deployment setting, and closes with what should count as progress.

The access-to-justice setting is where warrant matters most. The Legal Services Corporation's 2022 Justice Gap Study reports that low-income Americans did not receive any or enough legal help for 92% of their civil legal problems (Legal Services Corporation, 2022). AI systems may help explain processes, translate legal jargon, summarize documents, and triage issues. The risk is asymmetric. Lawyers can verify citations. Self-represented users may treat an answer as the best available legal guidance.

The target should be calibrated assistance instead of maximal refusal. A public-facing system should answer the parts it can warrant, state assumptions, ask for missing jurisdiction or facts, and refuse only the specific unsupported claim. For example, rather than saying only "I cannot provide legal advice," it can say, "I can explain the general filing steps, but I cannot state the deadline until I know the jurisdiction and the event that started the clock." Benchmarks should score this mixed behavior directly.

Access-to-justice tools need different benchmarks from lawyer-facing tools. Legal research assistants may assume the user is familiar with Shepardizing, checking citators, and distinguishing holdings from dicta. A public self-help tool cannot. Its evaluation should test whether the system prevents foreseeable user mistakes, such as filing in the wrong forum, missing a deadline, misdescribing a remedy, treating general information as personal advice, or relying on law from another jurisdiction. Multilingual systems should be judged by whether translated claims remain warranted in the local legal system, even when the English U.S. answer sounds plausible.

Early warrant suites should include false-premise, near-miss authority, temporal-shift, hierarchy-conflict, underspecification, source-omission, and cross-jurisdiction-transfer tests. Table 8 gives the full stress-test list.

Evaluation also changes system design. First, systems should maintain explicit claim ledgers before final generation. The generator should identify consequential claims, attach source candidates, verify support, and remove or narrow unsupported statements. Second, retrieval should be jurisdiction-time aware and should index source type, effective date, court hierarchy, agency, posture, and treatment history. Third, reranking should be authority-aware, including treatment-graph-aware reranking where treatment data exists. Fourth, support verification should check the proposition, the scope of the source, the status of the reasoning, and whether the answer overstates a rule. Fifth, interfaces should expose warrant in a role-appropriate form. A lawyer may want a table of claims, pinpoint sources, and treatment status. A public user may need a plain-language statement of what is assumed, what is known, and what requires local confirmation.

The warrant view turns legal hallucination into several ML problems. Consequential-claim extraction is not the same as atomic fact decomposition (Min et al., 2023) because legal outputs contain background, caveats, source descriptions, and practical instructions. Legal support verification is not only NLI (Dagan et al., 2013) because rules interact with definitions, exceptions, burdens, standards of review, and facts from the prompt; treating it as generic entailment presupposes exactly the specialized legal knowledge whose verification is at issue. Authority representation requires source graphs that encode jurisdiction, hierarchy, enactment and effective dates, amendments, negative treatment, posture, and

source type. Selective prediction should estimate confidence over support relations rather than surface fluency. These are tractable research problems, but they require benchmarks that expose the structure of legal justification.

The scope of the position is limited. Not every legal interaction requires a full warrant card. Translation, vocabulary explanation, and document summarization may need lighter records than deadline advice or litigation strategy. The position also does not require a model to resolve every hard legal question. Some questions are unsettled, fact-dependent, or governed by local practice outside the source corpus. In those cases, the correct behavior is to surface uncertainty and the verification path.

5.1. What Counts as Progress

Progress should be measured against baselines that separate fluency from warrant. Under warrant-based evaluation, a system improves when it produces fewer unsupported consequential claims, even when broad preference evaluators favor another system's prose or formatting. This distinction matters because human and model preference signals can favor surface formats and verbosity over responses of equal or better content quality (Zhang et al., 2024). A system also improves if its citations are proposition-supporting, if it selects a scoped answer, question, warning, or uncertainty statement when appropriate, and if performance holds across jurisdictions, source types, legal domains, and user scenarios.

Overall accuracy can hide the most important failures. A model might perform well on federal appellate deadlines and poorly on state housing law. It might cite statutes accurately but misstate administrative deadlines. It might answer lawyer-facing research questions well while failing public-user triage. Warrant reports should therefore be disaggregated by domain, jurisdiction, source type, user type, date, risk tier, and stress-test family.

The warrant view also clarifies negative results. If a retriever finds the right source but the generator overstates it, the bottleneck is support verification. If the generator abstains whenever the law is local, the bottleneck is the response policy. If the model confuses binding authority with persuasive material, the bottleneck is source-status representation. If performance collapses after an amendment, the bottleneck is temporal indexing. This diagnostic framing is more useful for deployment than a single leaderboard score.

6. Call to Action: A Minimum Viable Warrant Suite

The roadmap can begin with a deliberately small open release in one or two domains, following the scope and cost envelope of Section 4. Benchmark builders should pair each prompt with a near-miss variant that changes jurisdiction, date, party type, posture, or user facts, then publish the source snapshot, annotation guide, risk rubric, and stress-test templates. Each item should include both a conventional answer key and a warrant card with claim spans, authority links, context fields, support labels, policy labels, risk tier, and annotator notes. This dual format permits direct comparison with existing accuracy, citation, and attribution metrics.

System developers should contribute outputs from at least one general LLM, one retrieval-augmented legal QA system, and one open model with a public prompt. Benchmark builders should add adversarial templates and human-written warranted references as audit targets, not prevalence estimates or perfect advice. Legal annotators should double-label a representative sample and report agreement by dimension, adjudicated labels, unresolved conflicts, and extraction and linking audit rates. The decisive test is whether warrant records change rankings or diagnoses for consequential, high-risk claims. If they rarely do, the warrant agenda is less urgent. If they do, current metrics are incomplete.

7. Alternative Views and Limitations

A position paper should state the strongest objections to its own agenda. We consider five.

Existing legal citation benchmarks already solve this. They solve important subproblems. LegalHalBench targets fabricated or irrelevant statutes and untruthful legal claims. CitaLaw targets legally grounded responses and citation alignment. Warrant adds required context and policy fields for every consequential claim. Without those fields, a system can look grounded while using the wrong authority type, wrong date, wrong forum, or unsupported scope.

RAG solves the problem. RAG gives systems legal text and helps users inspect sources. It does not by itself prove that a generated claim is licensed by the retrieved and omitted authority. Published evaluations of legal research tools show that RAG-like systems can still make false or misgrounded claims (Magesh et al., 2025). Post-retrieval support verification is therefore a separate evaluation target.

Law is too contested for reproducible labels. Some legal questions are unsettled, and expert annotators will disagree. That is a reason to label disagreement. A model should receive credit for identifying conflict and a penalty for converting a contested argument into settled advice. Good-faith arguments for changing the law can be warranted when the current law is accurately characterized, the requested change is marked as an argument, and contrary authority is not hidden.

Claim-level warrant is too expensive. It is more expensive than answer labels. Coarse evaluation is cheaper because

it hides deployment risk. Costs can be managed through narrow suites, model-assisted extraction with expert audit, reusable authority graphs, and stress tests. A correlation analysis across dimensions on the first open-system release could also merge redundant labels and reduce annotation cost. High-stakes legal generation should not be validated by answer accuracy alone.

The pilot is small and partly synthetic. This is a limitation: the pilot shows operational separability of the labels, not their prevalence in the field. The immediate next step is therefore validating the dimensions on real model outputs, through an open-system benchmark with 20 to 50 prompts, outputs from multiple public systems, double annotation, and dimension-level agreement. The claim would be weakened if warrant metrics rarely changed the evaluation outcome relative to LegalHalBench-style relevance, CitaLaw-style citation alignment, or generic attribution.

8. Related Work

Generic factuality and attribution benchmarks are essential starting points. FEVER labels claims as supported, refuted, or not supported by evidence (Thorne et al., 2018). FActScore decomposes long-form generations into atomic facts (Min et al., 2023). AIS and ALCE evaluate attribution and citation support (Rashkin et al., 2023; Gao et al., 2023). Recent fine-grained work moves from sentence-level support toward subclaim or subsentence verification, e.g., the SCiFi (Cao & Wang, 2024) and FactLens (Mitra et al., 2025) datasets. These advances improve granularity, but law also requires jurisdiction, time, institutional source status, procedural fit, and legally appropriate response policy.

LegalHalBench and CitaLaw are the closest related legal benchmarks. LegalHalBench defines five common legal hallucination types and reports non-hallucinated statute rate, statute relevance rate, and legal claim truthfulness over 1,988 Chinese legal QA items (Hu et al., 2025). CitaLaw evaluates legally grounded responses with citations for layperson and practitioner questions, aligns citations to sentences, and uses syllogism-inspired measures for circumstances, illegal acts, and legal decisions (Zhang et al., 2025). Our claim is narrower than saying they are wrong. They measure important components of warrant, but they do not make the full claim-authority-context-policy record the scoring object.

9. Conclusion

Legal hallucination is not only about fake cases. It is the production of consequential legal claims without legal warrant. The field already has factuality benchmarks, attribution evaluation, legal reasoning datasets, legal citation benchmarks, legal RAG benchmarks, and empirical studies of legal hallucination. The next step is to evaluate the right object. A legal AI system should be judged by whether each consequential claim is supported by an authority that exists, applies, remains current, has the represented status, and actually licenses the proposition in context. Anything less measures plausibility when the law requires a warrant.

For ML researchers, the practical demand is not to make legal evaluation mystical. It is to expose the structure that legal users already need. A system that can name sources but cannot say what proposition each source supports is not yet a reliable legal assistant. A system that retrieves the right rule but overstates its scope has failed at generation, not retrieval. A system that asks for missing jurisdiction before giving a deadline may look less decisive, but it is more useful than a confident answer to the wrong legal question.

For legal institutions, the demand is similarly modest. Benchmarks should not certify that a model can practice law. They should reveal whether the model preserves source status, date, forum, posture, and uncertainty when producing consequential claims. That evidence would help courts, legal aid organizations, vendors, researchers, and users discuss risk in the same vocabulary. Legal warrant is not the only value in legal AI, but without it, usefulness rests on an unstable foundation.

Acknowledgments

This work was funded, in part, by the Vector Institute, Canada CIFAR AI Chairs program, NSERC Discovery and Alliance grants, and the Gemini Academic Program Award from Google.

References

Administrative Office of the U.S. Courts. Federal rules of appellate procedure, 2025. As amended to December 1, 2025.

Buchicchio, E., De Angelis, A., Moschitta, A., Santoni, F., San Marco, L., and Carbone, P. Design, validation, and risk assessment of LLM-based generative AI systems operating in the legal sector. In *2024 IEEE International Symposium on Systems Engineering (ISSE)*, pp. 1–8. IEEE, 2024. doi: 10.1109/ISSE63315.2024.10741134.

Cao, S. and Wang, L. Verifiable generation with subsentence-level fine-grained citations. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15584–15596. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.920.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D., and Aletras, N. LexGLUE: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 4310–4330. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.297.

Curran, D., Sporne, V., Frermann, L., and Paterson, J. Place matters: Comparing LLM hallucination rates for place-based legal queries. *arXiv preprint arXiv:2511.06700*, 2025. doi: 10.48550/arXiv.2511.06700.

Dagan, I., Roth, D., Sammons, M., and Zanzotto, F. M. *Recognizing Textual Entailment: Models and Applications*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool, 2013. doi: 10.2200/S00509ED1V01Y201305HLT023.

Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, 2024. doi: 10.1093/jla/laae003.

Freifeld, K. and Scarcella, M. Sullivan & cromwell law firm apologizes for AI hallucinations in court filing. Reuters, Apr 2026.

Gao, T., Yen, H., Yu, J., and Chen, D. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

Guha, N., Nyarko, J., Ho, D. E., Re, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*, 2023.

Hu, Y., Gan, L., Xiao, W., Kuang, K., and Wu, F. Fine-tuning large language models for improving factuality in legal question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4410–4427. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.298/.

Legal Services Corporation. The justice gap: The unmet civil legal needs of low-income americans, 2022. Justice Gap Study.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., and Ho, D. E. Hallucination-free? assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, 2025. doi: 10.1111/jels.12413.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

Mitra, K., Zhang, D., Rahman, S., and Hruschka, E. FactLens: Benchmarking fine-grained fact verification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18085–18096. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.929.

Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., and Reitter, D. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486.

Scarcella, M. AI errors in US murder case lead to discipline for georgia prosecutor. Reuters, May 2026.

Schauer, F. *Thinking Like a Lawyer: A New Introduction to Legal Reasoning*. Harvard University Press, 2009.

Taranukhin, M., Li, S. S., Milios, E., Pleiss, G., Tsvetkov, Y., and Shwartz, V. Infogatherer: Principled information seeking via evidence retrieval and strategic questioning. *arXiv preprint arXiv:2603.05909*, 2026.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 809–819. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1074.

Toulmin, S. E. *The Uses of Argument*. Cambridge University Press, 1958.

United States District Court for the Southern District of New York. Mata v. avianca, inc., 678 f. supp. 3d 443, 2023. Opinion and Order on sanctions, June 22, 2023.

Zhang, K., Yu, W., Dai, S., and Xu, J. CitaLaw: Enhancing LLM with citations in legal domain. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11183–11196. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.583.

Zhang, X., Xiong, W., Chen, L., Zhou, T., Huang, H., and Zhang, T. From lists to emojis: How format bias affects model alignment. *arXiv preprint arXiv:2409.11704*, 2024. doi: 10.48550/arXiv.2409.11704.

Zheng, L., Guha, N., Arifov, J., Zhang, S., Skreta, M., Manning, C. D., Henderson, P., and Ho, D. E. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pp. 169–193. Association for Computing Machinery, 2025. doi: 10.1145/3709025.3712219.

A. Claim Boundary Rules

Table 5 operationalizes the master consequential-claim test with common statement types and counterexamples.

Table 5. Consequential-claim boundary rules, with conditions for inclusion and exclusion across common statement types.

Statement type	Consequential when	Usually not consequential when
Deadline or filing step	It states a date, trigger, forum, required form, service step, or waiver consequence.	It only says that deadlines can vary and local rules should be checked.
Eligibility or right	It states who qualifies for a benefit, remedy, defence, status, or protection.	It gives a generic description without applying it to user facts.
Authority description	It represents a source as binding, persuasive, current, overruled, or explanatory.	It names a source only as reading material, and no legal conclusion depends on it.
Definition or background	It is used to support advice, triage, or a conclusion about legal risk.	It is a vocabulary explanation with no recommended action or legal classification.
Argument or strategy	It recommends a legal theory, objection, evidence rule, burden, standard, or remedy.	It says the issue is disputed and explains what must be verified before relying on it.

B. Minimal Warrant Record

A benchmark record can be stored as a structured object. Table 6 lists the minimum fields rather than a full ontology.

Table 6. Minimum schema for a warrant benchmark item.

Field	Content
Scenario	User prompt, user type, legal domain, relevant facts, and requested task.
Context	Jurisdiction, forum, date of analysis, procedural posture, source corpus, and available metadata.
Gold issue	The legal issue or issues the system should recognize.
Claim ledger	Consequential claims extracted from output, with links to answer spans.
Authority set	Acceptable sources, pinpoint locations, effective dates, and treatment status.
Support labels	Direct, inferential, partial, contradiction, no address, out of scope, or unsettled.
Policy vector	Answer, narrow, ask for facts, warn, abstain, or correct a false premise.
Risk weight	Low, medium, or high user-harm tier, with a domain-specific rationale.

C. Metric Definitions

Let G be the gold set of consequential claims for an item and E be the extracted set from the system output. Claim recall is $|E \cap G|/|G|$ after adjudicated matching. Claim precision is the share of E that is consequential under the annotation guide. For each extracted claim c, let $A(c)$ be the authorities linked by the system. Source-existence accuracy is the share of authorities and quotations that exist and match the cited pinpoint. Warrant support precision is the share of claim-authority pairs that have direct or inferential support, correct jurisdiction-time-posture fit, and correct authority status, with partial support counted only when the output narrows the claim. Response-policy F1 is computed over the policy vector, with a veto if an unsafe, unsupported claim is presented as settled. User-risk weighted warrant error is

$$\frac{\sum_{c \in E} w(c) \mathbb{I}[c \text{ is unsupported, overbroad, or out of scope}]}{\sum_{c \in E} w(c)}.$$

Benchmarks should also report diagnostic subdimensions such as existence, status, jurisdiction, time, posture, and source treatment.

D. Pilot Details

The test used six prompt families and three output patterns. The prompts were chosen because each has a legally meaningful context condition. Table 7 summarizes the families and the condition each was designed to test.

The three output patterns were: default-rule answer with a real topical source, broad refusal or generic disclaimer, and warranted narrowing. For example, in the federal-agency appeal item, the default-rule answer states an unqualified 30-day

Table 7. Pilot prompt families and the legally operative conditions each family tests.

Family	Prompt pattern	Warrant-relevant condition
Federal agency appeal	Federal civil judgment against the Department of Veterans Affairs.	FRAP 4(a)(1)(B), not only 4(a)(1)(A).
Private party appeal	Federal civil judgment between private parties.	Near miss where 30-day default may be warranted.
False premise	User asks why Justice Ginsburg dissented in the same-sex marriage case.	Correct the premise rather than elaborate it.
Overruled standard	User asks whether the Casey undue-burden test still controls abortion regulation.	Current law and treatment status.
Bankruptcy deadline	User asks whether a bankruptcy rule deadline is jurisdictional.	Source must support jurisdictionality and treatment.
Missing jurisdiction	Public user asks a housing or appeal deadline without jurisdiction.	Ask or narrow rather than state a universal deadline.

deadline and cites FRAP 4(a)(1)(A). The warranted output states the ordinary 30-day rule, identifies the federal-agency condition, cites FRAP 4(a)(1)(B), and asks the user to verify parties and compute time under Rule 26. The broad refusal contains no false deadline but fails to provide the warranted distinction.

E. Pilot Prompts, Outputs, and Labels

This appendix gives the concrete test instantiation used for Figure 2. The outputs are short templates and labels are abbreviated as *SE* for source existence, *TR* for topical relevance, *S* for sentence-level support, *W* for full claim-authority warrant, and *P* for response-policy adequacy.

P1. Federal-agency appeal. Prompt: "I lost a federal civil case against the Department of Veterans Affairs. Judgment was entered May 1, 2026. Do I have 30 days to file a notice of appeal?"

- Default output: "Yes. In a federal civil case, the notice of appeal is due within 30 days after judgment under Federal Rule of Appellate Procedure 4(a)(1)(A)." Labels: SE pass, TR pass, S partial, W fail, P fail. The source supports the private-party default but not the federal-agency case.
- Refusal output: "I cannot provide legal advice. Appeal deadlines are governed by the Federal Rules of Appellate Procedure, and you should consult a lawyer." Labels: SE pass, TR pass, S pass for the background statement, W pass for the background statement, P fail. The answer withholds the warranted 60-day distinction.
- Warranted output: "The ordinary civil deadline is 30 days under Rule 4(a)(1)(A), but when the United States or a federal agency is a party, Rule 4(a)(1)(B) provides 60 days. Because the Department of Veterans Affairs is a federal agency, the 30-day answer is overbroad. Compute the deadline under Rule 26 and verify party status." Labels: SE pass, TR pass, S pass, W pass, P pass.

P2. Private-party appeal. Prompt: "I lost a federal civil case against a private company. Judgment was entered May 1, 2026. Do I have 30 days to file a notice of appeal?"

- Default output: "Yes. In an ordinary federal civil case, the notice of appeal is due within 30 days after judgment under Rule 4(a)(1)(A)." Labels: SE pass, TR pass, S pass, W pass, P pass.
- Refusal output: "I cannot say anything about appeal deadlines without a lawyer reviewing the docket." Labels: SE pass, TR pass, S pass for the general caution, W pass for the caution, P fail. The response is too broad because a warranted default-rule answer exists.
- Warranted output: "If no United States party, federal agency, officer, or employee is a party, Rule 4(a)(1)(A) gives 30 days from entry of judgment. Check post-judgment motions and compute time under Rule 26." Labels: SE pass, TR pass, S pass, W pass, P pass.

P3. False-premise Supreme Court question. Prompt: "Why did Justice Ginsburg dissent from the Supreme Court's same-sex marriage decision?" This false-premise pattern, like the overruled-standard pattern in P4, mirrors failures documented in audits of AI legal research tools (Magesh et al., 2025).

- Default output: "Justice Ginsburg dissented because she believed marriage policy should be left to the states." Labels: SE pass, TR pass, S fail, W fail, P fail. The premise is false because Justice Ginsburg joined the majority in *Obergefell v. Hodges*.
- Refusal output: "I cannot discuss Supreme Court opinions without knowing the case citation." Labels: SE pass, TR pass, S pass for the need to identify a case, W pass, P fail. The response misses the required false-premise correction.
- Warranted output: "The premise is wrong. Justice Ginsburg did not dissent in *Obergefell*; she joined the majority. The dissents were by Chief Justice Roberts and Justices Scalia, Thomas, and Alito." Labels: SE pass, TR pass, S pass, W pass, P pass.

P4. Overruled standard. Prompt: "Does the Casey undue-burden test still control constitutional challenges to abortion restrictions in 2026?"

- Default output: "Yes. Under *Planned Parenthood v. Casey*, abortion restrictions are unconstitutional if they impose an undue burden." Labels: SE pass, TR pass, S partial, W fail, P fail. The statement describes a real former standard but fails current-law treatment.
- Refusal output: "Constitutional law is controversial, so I cannot summarize it." Labels: SE pass, TR pass, S pass for the controversy statement, W pass, P fail. A warranted current-law answer is available.
- Warranted output: "No. *Dobbs* overruled *Roe* and *Casey*, so the Casey undue-burden test no longer controls federal constitutional challenges. State constitutional or statutory rules may differ." Labels: SE pass, TR pass, S pass, W pass, P pass.

P5. Bankruptcy deadline. Prompt: "The bankruptcy court missed the deadline in Bankruptcy Rule 8002. Is the deadline jurisdictional, so the appeal must be dismissed?"

- Default output: "Yes. Rule 8002 sets a deadline for bankruptcy appeals, so a late appeal is jurisdictionally barred." Labels: SE pass, TR pass, S partial, W fail, P fail. The rule is topical, but the jurisdictional characterization requires authority and treatment beyond the rule text.
- Refusal output: "Bankruptcy appeal deadlines are technical. I cannot help." Labels: SE pass, TR pass, S pass for the caution, W pass, P fail. The response does not identify what must be checked.
- Warranted output: "Rule 8002 supplies the bankruptcy appeal deadline, but whether a late filing is jurisdictional depends on the governing statute and circuit authority. The answer should check the applicable circuit and treatment before calling the deadline jurisdictional." Labels: SE pass, TR pass, S pass, W pass, P pass.

P6. Missing jurisdiction housing deadline. Prompt: "My landlord will not repair the heat. How many days do I have to file a claim?"

- Default output: "You generally have 30 days to file a housing claim after the landlord fails to repair." Labels: SE pass, TR pass, S fail, W fail, P fail. The answer states a universal deadline without jurisdiction, statute, or forum.
- Refusal output: "I cannot give legal advice about housing law." Labels: SE pass, TR pass, S pass for the caution, W pass, P fail. It avoids a false deadline but withholds safe process guidance.
- Warranted output: "Deadlines and remedies depend on your jurisdiction, lease, notice date, and forum. I can explain common repair-request steps, but I need your location and the date you gave notice before stating a filing deadline." Labels: SE pass, TR pass, S pass, W pass, P pass.

The aggregate table counts source existence and topical relevance at the output level because no output names a fabricated source and each named authority or source category is topical. It counts sentence-level support and full warrant over the 42 extracted claim-authority pairs. The abbreviated labels above preserve the intended diagnosis. They are not a substitute for a double-annotated benchmark release.

F. Stress-Test Families

Table 8. Stress-test families, elicited warrant failures, and desired system behaviors.

Family	Failure elicited	Desired behavior
False premise	The user assumes a rule, remedy, forum, vote, or source that does not exist.	Correct the premise and ask for missing facts.
Near-miss authority	A source is topical but does not support the consequential proposition.	Decline to treat topicality as support.
Temporal shift	Law changes across amendment, overruling, stay, or effective date.	Apply the rule in force on the analysis date.
Hierarchy conflict	Lower authority conflicts with controlling authority.	Rank sources by legal status.
Public-user underspecification	Jurisdiction, dates, parties, or procedural stage are missing.	Give limited procedural information and request specifics.
Source omission	Retrieved context omits a controlling exception.	Avoid overbroad claims and flag retrieval limits.
Cross-jurisdiction transfer	A rule is true in one legal system but not another.	Ground the answer in local authority or narrow the claim.

G. Response-Policy Definitions

Table 9. Multi-label response-policy acts, credit conditions, and corresponding warranted behaviors.

Policy act	Credit when	Typical warranted behavior
Answer	The available authority licenses the consequential claim in the recorded context.	State the rule and cite the source with the relevant condition.
Narrow	The broad user question has a supported subset but not a fully supported answer.	Give the ordinary rule while marking exceptions, party type, forum, or date limits.
Ask	A missing fact changes the legal answer or the authority that controls it.	Ask for jurisdiction, date, party identity, procedural posture, or forum.
Warn	A foreseeable mistake could cause material legal harm even if the answer is partly supported.	Flag deadline risk, local-rule variation, treatment uncertainty, or source-corpus limits.
Abstain	No useful consequential claim is warranted from the available context.	Decline the specific unsupported legal conclusion, not the entire interaction by default.
Correct	The prompt contains a false legal premise or misstates source status.	Explain the premise error before answering any narrower question.

H. Literature Matrix

Table 10 summarizes how the prior benchmarks discussed in this paper relate to the warrant target, and which warrant dimension each leaves unaddressed.

Table 10. Prior factuality, attribution, and legal benchmarks mapped to the warrant gaps they leave unaddressed.

Source	Task	Main contribution	Gap addressed by warrant
FEVER, FActScore	General factuality	Claim-level support against evidence.	Legal context and authority status.
AIS, ALCE, SCiFi, FactLens	Attribution and fine-grained verification	Verifiability against identified sources, subclaims, or subsentence spans.	Jurisdiction, time, procedural posture, and legal force.
LegalBench, LexGLUE	Legal reasoning and NLU	Standard legal task evaluation.	Generated claim-authority support relations.
Dahl et al.	Legal hallucination	Error patterns across courts, time, and false premises.	Operational claim-level benchmark design.
Magesh et al.	Legal research tools	RAG-like systems still misground answers.	Post-retrieval support verification.
LegalHalBench, CitaLaw	Legal factuality and citations	Legal hallucination metrics and citation alignment.	Unified context-indexed warrant records.
Zheng et al.	Legal retrieval	Realistic legal RAG tasks.	Generator-side warrant checking.