Grounding Measurement for Legal AI

A methodology for independent audit against the court-adjudicated record

Ratia

Minnesota · 2026

Contents

This paper describes the methodology by which Ratia measures the grounding of legal-AI output against the court-adjudicated record. It is intended for firms, insurance carriers, vendors, and regulators evaluating what "accurate" means when a legal tool cites case law. It is not legal advice.

Executive summary

Legal-AI vendors report accuracy rates in the high nineties. Courts, meanwhile, have addressed AI errors in more than fourteen hundred filings. The gap between what vendors measure and what courts sanction is the subject of this paper.

The gap is not an accident of adoption. It is a measurement problem. Every legal-AI vendor grades its own work. No two vendors grade the same way. None of the reported numbers are anchored to what a court actually held. Buyers most exposed to the downside — law firms, and the malpractice carriers that underwrite them — are choosing between tools on the basis of vendor demos.

Ratia is an independent audit of legal-AI output. It measures three things, each against the court-adjudicated record itself, at the level of the pincite:

1. **Exists.** The cited case is real, and the opinion says what the tool claims it says.
2. **Holds.** The tool has represented the holding on the dispositive issue correctly.
3. **Valid.** The case is still good law — not reversed, overruled, or superseded.

Each check is defined in this paper, operationalized in a rubric, applied blind by two independent attorney reviewers, adjudicated on disagreement, and reported by court level and issue type. Nothing is blended into a flattering composite. The methodology is documented so that it can be inspected, criticized, and replicated.

The Ratia Report is the output. It is designed to be read by an underwriter, a partner, an ethics panel, and an expert witness. It says what the tool got right and what it got wrong; it does not tell the reader whether to buy the tool.

The pilot jurisdiction is Minnesota. This paper explains why we started there, how we constructed the ground-truth corpus for Minnesota state law, and what expansion to other jurisdictions will require.

The paper closes with a section on limits. Ratia audits are snapshots — of a specific tool version, on a specific jurisdictional corpus, at a specific date. They are not benchmarks. They do not evaluate every downstream use of AI in legal practice. They are not certifications, and they are not legal advice.

---

1. The measurement gap

1.1 What vendors report

Every product marketed as “legal AI” reports an accuracy figure. The figures typically live somewhere between 95% and 99%. They are calculated on internal test sets, using rubrics defined by the vendor, evaluated by reviewers selected and compensated by the vendor. The methodology is rarely published. The test sets are almost never published. Where a rubric is disclosed at all, it typically counts as “accurate” any answer whose surface features match a

training-set expectation, without regard to whether a court adjudicated the underlying question.

This is not fraud. It is a standard state of affairs in a young industry that has not yet been made to disclose. It is also not measurement, in the sense that a firm buying the tool, or a carrier underwriting the firm, would understand measurement. The vendor's number and the buyer's exposure are not on the same axis.

1.2 What courts have seen

Judges in state and federal courts have addressed AI errors in filings in more than fourteen hundred matters. The reported patterns:

- **Non-existent citations.** Cases that do not exist, cited as if they do. In some instances, entire fabricated case names with plausible-looking reporters and page numbers. In others, real party names attached to nonexistent opinions.
- **Misread holdings.** Real cases cited for propositions the opinion does not actually support. The case is real; the holding is not what the tool claims.
- **Dead law.** Cases that exist and once said what the tool claims, but were reversed, overruled, or superseded. The tool did not check.
- **Wrong jurisdiction.** Tools not designed for the practice jurisdiction returning authorities from a different state's law as if they were controlling.

Every one of these categories has been named in a sanctions order. Every one of these categories is invisible to the vendor's self-reported accuracy number.

1.3 Why the gap persists

The gap persists for reasons that are structural, not moral. Three of them:

Grading your own work. No independent body currently measures legal-AI grounding. Third-party benchmarks that exist (legal-domain LLM evaluations from academic groups, some vendor-comparison studies) measure narrow tasks — bar exam questions, contract classification — that are not the same task as "cite the controlling authority on a Minnesota holding-water dispute." A vendor cannot be graded down for a category no one is grading in.

Fluency vs. grounding. Modern general-purpose language models produce prose that reads as competent legal writing. Fluency is a design goal of the underlying model. Grounding is not. A tool that has been prompted to write like a lawyer will do so — with citations, with the right voice, with the syntax of a memorandum — even when the citations do not resolve to real opinions. Buyers who evaluate on the read-through cannot tell fluent from grounded, because both look the same on the page.

The Claude-wrapper reality. A large fraction of what is marketed as legal AI is a general-purpose model built by Anthropic, OpenAI, or Google — Claude, ChatGPT, Gemini — with a legal-themed system prompt, sometimes a retrieval layer, and a slicker interface. Nothing

wrong with that as a construction. But it means the AI your firm is running is, in many cases, the same underlying model your opposing counsel is running through a different-colored product. The model is not what differentiates. The prompt, the retrieval, the guardrails, and the evaluation are. Of those four, only the first three are ever discussed in marketing. The fourth — evaluation — is almost never disclosed, and never measured independently.

1.4 What Ratia does

Ratia is an independent audit. We take a legal-AI tool's actual output — real citations, real characterizations of holdings, real claims about current validity — and measure it against the court-adjudicated record itself. We do not measure against a headnote, a summary, an editorial digest, or a training-set proxy. We measure against the opinion the court wrote, at the pincite the tool claims to be citing.

The remainder of this paper describes how.

---

2. Definitions

The single largest source of confusion in this area is that the words carry different meanings for different audiences. Before describing method, it is worth being precise.

2.1 Grounding

In machine-learning research, "grounding" refers to the connection between a model's output and some external referent — training data, a retrieval source, a knowledge base. A model output is "grounded" if it can be traced back to a source; "hallucinated" if it cannot.

For legal AI, that generic definition is insufficient. A tool can be grounded in a training corpus and still be wrong about what a court held. A tool can be grounded in a retrieved document (a Westlaw headnote, an editor's summary, a treatise) and still misrepresent the opinion the retrieved document was describing. The relevant question for the profession is not whether the output has *any* source, but whether it has the *right* source: the court's own words, at the pincite, on the issue being cited for.

For purposes of this methodology, **grounding** is defined narrowly:

> A legal-AI output is *grounded* in an authority if, and only if, the exact claim the output makes about that authority is supported by the text of the court's opinion at the pincite cited.

"Supported by" means: an attorney reading the opinion at the pincite, without prompting from the tool, would characterize the holding or proposition the same way the tool did.

This is a strong definition. It excludes headnote-only grounding, editorial-summary grounding, and treatise-derived grounding. Those forms of grounding may be perfectly acceptable in litigation practice — attorneys use headnotes and summaries constantly. But they are not

the same thing as pincite-grounded, and the distinction matters for the audit, because the sanctions in the reported cases have landed where the tool's output diverged from the *opinion*, not from the *summary*.

2.2 Hallucination

"Hallucination" in the AI literature refers to any output for which the model has no supporting source. In legal context, we use it more narrowly:

> A legal-AI output is a *hallucination* if it makes a claim about an authority that cannot be verified against the court-adjudicated record — because the authority does not exist, because the authority exists but the opinion does not support the claim, or because the opinion supported the claim once but no longer does.

Hallucinations, so defined, fall into three subtypes that correspond to the three checks we run: existence hallucinations (Section 3.1), holding hallucinations (Section 3.2), and validity hallucinations (Section 3.3).

2.3 The court-adjudicated record

The **court-adjudicated record** is the set of judicial opinions issued by courts of the relevant jurisdiction, in their published or authoritative form, together with the subsequent history of each opinion (affirmances, reversals, denials of review, overrulings). It is not the set of headnotes. It is not the set of editor summaries. It is not the set of secondary sources.

The court-adjudicated record is what a court would take judicial notice of. It is what a lawyer signing a filing is representing that lawyer has read. It is what the sanctioned attorneys in the reported cases were, in effect, sanctioned for not consulting.

For the Minnesota pilot, the court-adjudicated record consists of:

- Minnesota Supreme Court opinions (published)
- Minnesota Court of Appeals opinions (published and unpublished, with the unpublished status flagged)
- Federal circuit and district opinions applying Minnesota law
- Subsequent-history data (denials of review, superseding decisions, statutory abrogations affecting the holding)

2.4 Pincite

A **pincite** is a specific page (or paragraph in some reporter styles) within an opinion where the cited proposition appears. A citation without a pincite locates the opinion but not the point being made. A citation with an incorrect pincite locates a page of the opinion that does not say what the tool claims.

Pincite-level analysis is central to this methodology because that is the level at which the grounding claim is falsifiable. "The case says X somewhere" is not falsifiable in a useful

sense. “The case says X at page 217” is either true or not.

2.5 Holding, dictum, editorial summary

Distinctions the audit turns on:

- A **holding** is a proposition of law that was necessary to the court’s decision on the issue before it.
- A **dictum** is a proposition of law stated in the opinion but not necessary to the decision. Dicta bind no one, but tools frequently cite them as holdings.
- An **editorial summary** is a description of the opinion prepared by a publisher (West, Lexis, and others) after the fact. Editorial summaries are useful; they are not the opinion, and the court did not write them.

A tool that quotes a headnote as if it were the court’s own words is committing a distinct category of error — not always a sanctionable one, but relevant to the audit.

---

3. The three checks

Ratia audits a tool’s output on three checks. Each check is scored independently. Nothing is blended into a composite.

3.1 Exists — the case is real

Question. Does the cited case exist as a real judicial opinion, and does the opinion, at the cited pincite, say something recognizably related to what the tool claims?

Failure modes. - The case does not exist. The tool has fabricated a case name, a reporter, a page number, or all three. - The case exists but the pincite does not exist (page 217 of an opinion that ends at page 195). - The case exists and the pincite is real, but the pincite does not say what the tool claims. This is the border between an existence failure and a holding failure; we classify it as an existence failure when the pincite text is unrelated to the tool’s claim, and as a holding failure when the pincite text is on-topic but characterized incorrectly.

Method. 1. Parse the tool’s citation into components: party names, reporter volume, reporter page, court, year, pincite. 2. Resolve against the jurisdiction’s authoritative case-law corpus. 3. Retrieve the opinion. 4. Locate the pincite. 5. Extract the pincite text (paragraph or page). 6. Compare pincite text to the tool’s claim for topical relatedness.

Scoring. Pass, fail, or partial. Partial is reserved for cases where the case exists and the pincite exists but the citation format is materially malformed in a way that would be flagged by a court but does not defeat verification (e.g., wrong reporter series, correct volume and page).

Illustrative example (composite, not real). A tool cites `Anderson v. Duncan Financial, 823 N.W.2d 445, 452 (Minn. Ct. App. 2019)` for the proposition that a lender must provide

a payoff statement within ten days of a written request. We look up 823 N.W.2d. Volume 823 does not include a Minnesota Court of Appeals opinion at page 445; the volume includes a Wisconsin Supreme Court opinion at page 439 and a Michigan Court of Appeals opinion at page 460. There is no such case. Existence fail.

3.2 Holds — the holding was read correctly

Question. Given that the case exists and the pincite is on-topic, did the tool characterize the holding correctly on the issue for which it is cited?

Failure modes. - Holding overreach: the tool states a broader proposition than the opinion supports. - Holding narrowing: the tool states a proposition narrower than the opinion actually supports (rarer, still an error). - Holding-dictum confusion: the tool cites dictum as if it were a holding. - Wrong issue: the opinion addresses two issues; the tool assigns the holding on one to the other. - Headnote-parroting: the tool quotes an editorial headnote as if it were the court's language, when the opinion says something materially different at the pincite.

Method. 1. Two attorney reviewers, each independently, read the opinion (not the tool's characterization). 2. Each reviewer, blind to the other and to the tool's identity, characterizes the holding on the relevant issue in their own words. 3. Each reviewer, then, compares the tool's characterization to their own characterization on a defined rubric (Appendix A). 4. Reviewer scores are compared. Agreement produces a final score; disagreement triggers adjudication by a third reviewer.

Scoring. Pass, fail, or partial. Partial is reserved for cases where the tool's characterization is correct in substance but overreaches or narrows in a way that could mislead a reasonable attorney reader.

Illustrative example (composite, not real). A tool cites a Minnesota Supreme Court opinion for the proposition that "a party's failure to raise an issue on appeal waives it." The opinion in fact holds that failure to raise an issue *in a brief on appeal, where the issue was raised in the trial court*, waives it *for purposes of that appeal but not for post-conviction relief*. The tool's version is a substantial overreach of the actual holding, in a way that would matter to a client. Holding fail.

3.3 Valid — it is still good law

Question. Given that the case exists and the holding was read correctly, is the case still good law on the point cited?

Failure modes. - **Reversed on appeal.** The cited authority was reversed by a higher court and the tool did not check. - **Overruled.** A later opinion in the same court explicitly overruled the cited opinion, in whole or on the point cited. The tool did not check. - **Superseded by statute.** The legislature has enacted a statute that displaces the common-law rule the opinion articulated. The tool did not check. - **Distinguished into irrelevance.** The rule survives in name but has been distinguished by subsequent decisions to the point that it no longer

supports the proposition for which the tool cites it.

Method. 1. For each cited authority, retrieve the subsequent-history data. 2. Read any citing opinions that (a) reversed, overruled, superseded, or (b) are flagged by editorial services as negative treatment. 3. An attorney reviewer determines whether the case remains authoritative on the point the tool cited it for.

Scoring. Pass, fail, or partial. Partial is reserved for cases where the authority remains valid on some grounds but has been meaningfully limited on the point cited.

Illustrative example (composite, not real). A tool cites a 2011 Minnesota Court of Appeals opinion for a proposition about the scope of a landlord's duty to maintain habitability. In 2018, the Minnesota Supreme Court, in an opinion the tool does not cite, expressly rejected the Court of Appeals reasoning and articulated a different rule. The 2011 opinion has not been overruled by name; the Supreme Court's later opinion controls. Validity fail.

---

4. Ground-truth construction

The audit is only as credible as the ground-truth it measures against. Ground truth is not an assumption; it is a construction, and its construction is where the methodology succeeds or fails.

4.1 Sources

For the Minnesota pilot, the ground-truth corpus is built from:

- The Minnesota Supreme Court's official opinion PDFs, as published on the court's website, from 1858 to date.
- The Minnesota Court of Appeals' official opinion PDFs, published and unpublished, from 1983 (the court's inception) to date.
- Federal circuit and district opinions applying Minnesota law, drawn from the Federal Reporter and Federal Supplement series.
- Subsequent-history data derived from a combination of the courts' own records, editorial-service flagging (which we cross-check against the opinions themselves), and manual review of citation trails for opinions that have been substantively criticized.

Where possible, we anchor to the court's own official records rather than to editorially-processed versions. Where an official version is not available (older unpublished Court of Appeals decisions, for example), we use the editorial version and note the source.

4.2 Structure

The corpus is structured for pincite-level retrieval. Each opinion is stored with:

- Full opinion text with preserved paragraph and page numbers.
- Structured metadata: court, panel, date, disposition, judges.

- Subsequent-history index: reversed, affirmed, denied review, cited by, distinguished by, overruled by (each with source citation and date).
- Statutory-abrogation index: statutes enacted after the opinion that may affect its precedential weight on identified issues.

The pincite structure allows for retrieval at the level actually cited. A tool that cites "823 N.W.2d 445, 452" is claiming its proposition appears on page 452 of that reporter. Ground truth for that claim is the text on page 452, not the case as a whole.

4.3 Coverage limits

The Minnesota pilot corpus is complete for:

- Minnesota state appellate opinions from 1983 forward
- Minnesota Supreme Court opinions from 1858 forward
- Federal opinions applying Minnesota law from 1990 forward

The corpus is incomplete for:

- Trial-court decisions (Minnesota district court decisions are not systematically published; where a tool cites one, we verify what we can and flag limits)
- Bankruptcy court decisions applying Minnesota law (partial coverage)
- Older federal opinions (1990 backward, partial coverage)

The audit report for any given tool specifies which cited authorities fell within the fully-covered corpus and which fell outside. Authorities outside the covered corpus are not scored; they are noted as unscored, with the reason.

4.4 Why Minnesota first

Three reasons.

Scale. Minnesota's appellate output is on the order of 400-600 opinions per year across the Supreme Court and Court of Appeals. This is small enough to permit complete pincite-level indexing by a small team, and large enough to represent a realistic operating jurisdiction. Federal law and multi-state comparisons compound in ways that would defeat a startup-scale audit.

Coherence. Minnesota's appellate case law is well-organized and centrally published. The courts' own records are available in structured form. Editorial-service coverage is complete. This reduces the ambiguity in ground-truth construction.

Familiarity. The people building Ratia know Minnesota law. That is not a scalable advantage, but it is a real one at the pilot stage, where the risk of misreading a holding is highest.

Expansion to additional jurisdictions is discussed in Section 8.

---

5. Scoring

5.1 Blindness

To reduce evaluator bias, the audit is conducted blind at two levels:

- **Vendor blind.** Reviewers do not know which tool produced the output they are evaluating. The tool's output is stripped of identifying markers before it reaches the reviewer. The reviewer sees only: a citation, a proposition-claim, and (where applicable) any quoted text the tool attributes to the opinion.
- **Cross-reviewer blind.** Each reviewer is blind to the other reviewer's score until adjudication.

Blindness is a discipline, not a claim about human nature. Reviewers are attorneys with reputations to maintain. But even careful reviewers unconsciously calibrate to expectations. Blindness closes that channel.

5.2 Rubric

Each check (Exists, Holds, Valid) is scored on a defined rubric (Appendix A). Rubric application is documented per-authority: what the tool claimed, what the record showed, and which rubric level applied. This documentation is what makes the score inspectable.

5.3 Inter-rater reliability

Every audit reports the Cohen's kappa between the two blind reviewers on each check, computed across the audited output. This is the first-order signal for whether the rubric is well-defined and applied consistently.

Kappa targets, pre-adjudication:

- Exists check: $\kappa \geq 0.85$ (the check is largely mechanical once the corpus is complete)
- Holds check: $\kappa \geq 0.70$ (the check involves judgment; substantial agreement is the target)
- Valid check: $\kappa \geq 0.80$

Where kappa falls below target on any check, the audit report includes a discussion of where reviewers diverged and how the rubric will be tightened in the next audit cycle.

5.4 Adjudication

Any disagreement between the two blind reviewers is adjudicated by a third reviewer, senior to both, who reviews both scoring documents and the underlying opinion. The adjudicator writes a brief opinion on the disagreement and enters a final score. Adjudication rates are reported.

5.5 What is not scored

- The tool's fluency, tone, or writing quality

- The tool’s coverage (whether it retrieved authorities that a good attorney would have retrieved but the tool missed) — this is a related but distinct check, called retrieval-recall, and is out of scope for a grounding audit
- The tool’s response to prompt variations
- The tool’s compliance with any bar rule or ethical standard

These are all interesting questions. They are not questions Ratia claims to answer. Confining scope is part of what makes the score defensible.

---

6. The Ratia Report

The output of an audit is a document — the Ratia Report — sized to be read in one sitting by an underwriter, a partner, an ethics panel, or an expert witness in a subsequent proceeding.

6.1 Structure

Every Ratia Report contains:

Scope statement. Which tool, which version, which jurisdiction, which time period of tool output, which subset of cited authorities was in the ground-truth-covered corpus, which was out.

Scorecard. For each check (Exists, Holds, Valid), the pass / partial / fail counts and rates, reported by court level (state Supreme Court, state Court of Appeals, federal courts applying state law) and by output category (memorandum, brief, research note, etc., where the audit examined more than one).

Failure profile. For each failure, a short description of what the tool claimed and what the record showed. Failures are grouped by type (existence: fabricated case; holding: overreach; validity: superseded by statute; and so on). Frequency of each type is reported.

Sample outputs. Illustrative failures with the tool’s original output and the corresponding record, annotated for the reader.

Kappa and adjudication. Inter-rater reliability numbers, adjudication rate, and discussion of any check that fell below target.

Methodology reference. A reference to this methodology paper, with the version number of the rubric applied.

Limits. What the audit did not test. What the reader should not infer.

6.2 What the Report is not

The Ratia Report is not a certification. It does not certify a tool for general use. It does not say a tool is safe. It does not say a tool complies with any bar rule.

The Ratia Report is not a benchmark. Different audits are not directly comparable unless they

were conducted on the same rubric version, the same jurisdictional corpus, and the same output categories. Where prospects want cross-audit comparisons, we conduct dedicated comparative studies and report them separately.

The Ratia Report is not legal advice. It is a measurement of the grounding of specified outputs. What a firm should do about that measurement — how to weight it against other considerations, how to communicate it to clients, how to reflect it in engagement letters — is a decision the firm makes, with its own counsel.

6.3 Distribution

Reports are prepared for the party that commissioned the audit. That party controls distribution. Ratia does not publish audit results without the commissioning party's consent, and does not disclose the identity of parties whose audits are private.

Aggregate research — patterns across the audits we have conducted, without identifying any particular tool — may be published as separate research notes, with all identifying detail removed.

---

7. Limits and honest caveats

The audit is a serious instrument. It is not an omniscient one. The following are the limits the methodology imposes on itself.

7.1 A snapshot in time

Legal-AI tools change. A model is updated. A retrieval layer is expanded. A prompt is tuned. A vendor releases a new version. Any audit result is dated: to the tool version, the rubric version, the jurisdiction, and the calendar date on which the audit was conducted. A report from six months ago is not a claim about the tool today.

7.2 Jurisdiction-bounded

The Minnesota pilot corpus supports audits of Minnesota-law claims. It does not support audits of a tool's output on New York law, Delaware corporate law, or federal circuit-court doctrine outside the Eighth Circuit. Expansion to other jurisdictions is a corpus-construction problem, not a methodological one; the check definitions and scoring transfer, but the ground-truth data does not.

7.3 Grounding is not the whole story

A perfectly grounded tool can still be a bad tool. It might retrieve the wrong authorities. It might miss authorities a competent attorney would have found. It might strategize badly. It might communicate confusingly. Grounding is a necessary condition for a legal-AI tool to be responsibly deployed; it is not a sufficient one.

Ratia audits grounding. Other axes of tool quality — retrieval recall, argumentation, drafting, tone — are separate research problems.

7.4 Human review is human

Attorney review of holdings involves judgment. Reasonable attorneys will sometimes disagree on how to characterize a holding. The rubric constrains that disagreement, and inter-rater reliability quantifies it, but does not eliminate it. The methodology is honest about this. Where kappa is below target, we say so.

7.5 Not a certification, not advice

Bearing repeating. A Ratia Report is a measurement. It is not:

- A certification that a tool is safe for a particular use
- Legal advice about any bar rule or ethical duty
- An opinion on any specific client matter
- A guarantee of the tool's future behavior

7.6 Ratia is not immune to error

We make mistakes. When a report contains a scoring error, we correct it, note the correction in a public erratum, and disclose the correction to the party that commissioned the audit and to any downstream party that received the report through official channels. We track our error rate and report it in aggregate.

---

8. Roadmap

8.1 Pilot to production

The Minnesota pilot is the first application of this methodology. Its purpose is to prove three things:

1. That the corpus can be constructed to the standard required for pincite-level verification.
2. That the rubric can be applied by attorney reviewers at inter-rater reliability targets.
3. That the resulting report is useful — that firms, carriers, and vendors read it, understand it, and make different decisions on the basis of it.

If those three things prove out, the methodology moves from pilot to production and the audit becomes a repeatable service.

8.2 Second jurisdictions

Corpus construction for a second jurisdiction is a defined project. It requires:

- Identifying the authoritative sources for that jurisdiction's appellate opinions

- Ingesting and structuring the opinions at pincite level
- Building subsequent-history data
- Recruiting attorney reviewers with jurisdictional expertise
- Adapting the rubric to any jurisdictional peculiarities (some jurisdictions have distinctive treatment of unpublished opinions, for example)

Rough sizing: three to nine months per jurisdiction, depending on the size of the appellate corpus and the availability of structured sources. Expansion priorities will follow demand.

8.3 Comparative studies

A single-tool audit answers "how does this tool perform on this jurisdiction." A comparative audit answers "how do these several tools compare on this jurisdiction." Comparative studies are more expensive to conduct — the rubric must be applied to every tool's output in the same window, with the same reviewers — but they answer the question buyers actually have. Comparative studies will be undertaken as jurisdictional corpora reach production stability and as vendor cooperation, or lack of it, permits.

8.4 Public research

Where audit patterns are large enough to draw general conclusions, and where identifying detail can be scrubbed, we intend to publish aggregate research. Sample topics on the roadmap:

- Failure-mode distribution across audited tools
- Effect of tool version updates on grounding rates
- Relationship between vendor-reported accuracy and audited grounding
- Effect of retrieval-layer design on grounding rates
- Court-level variation in grounding failure (are appellate-court cites more reliably grounded than trial-court cites?)

Publications will be Ratia research notes, cited to this methodology paper for procedure.

Appendix A — Rubric outline

(This appendix will be expanded into a full rubric document in a subsequent publication. What follows is the operational skeleton.)

Exists check

Level	Description
Pass	Cited case exists at the cited reporter and page; pincite is on-topic.

Level	Description
Partial	Case exists; citation format is malformed but resolvable (e.g., wrong series, correct volume/page).
Fail — Fabricated	No such case at that reporter and page.
Fail — Pincite	Case exists; pincite does not exist within the opinion.
Fail — Unrelated	Case exists; pincite exists; pincite text has no relation to the tool's proposition.

Holds check

Level	Description
Pass	Tool's characterization of the holding matches the record on the dispositive issue.
Partial — Overreach	Tool's characterization is directionally correct but broader than the opinion supports.
Partial — Narrowing	Tool's characterization is directionally correct but narrower than the opinion supports.
Fail — Dictum	Tool cites dictum as if it were a holding on the issue.
Fail — Wrong issue	Tool ascribes to one issue a holding the court reached on a different issue.
Fail — Headnote	Tool quotes editorial headnote as if it were the court's language, and the opinion says something materially different.
Fail — Fabricated	Tool's characterization does not appear in the opinion at any point.

Valid check

Level	Description
Pass	Authority remains good law on the point cited.
Partial — Limited	Authority remains good law generally but has been limited on the specific point cited.
Fail — Reversed	Authority was reversed on direct appeal; tool did not note.
Fail — Overruled	Authority was expressly overruled by a later opinion of the same court on the point cited; tool did not note.
Fail — Superseded	Authority was displaced by statute on the point cited; tool did not note.
Fail — Abrogated	Authority was distinguished into irrelevance by subsequent decisions on the point cited; tool did not note.

Appendix B — Worked example

This appendix walks a single tool output through the full audit process — retrieval, rubric application, blind scoring, adjudication, and reporting. The case, the citation, the tool output, and the reviewer notes are all constructed for illustration. Nothing in this example is a real audit result and no real vendor is implicated. It is included so that readers can see how the methodology would run on a real audit.

B.1 The audit setup

An unnamed vendor's legal-AI research assistant is under audit for a Minnesota employment-law engagement. The audit period covers thirty days of tool outputs on real research queries submitted by attorneys at three participating firms. This appendix shows the audit of one output.

B.2 The tool output

A firm associate asked the tool: *"What is the standard for a Minnesota at-will employer to terminate an employee, and are there exceptions?"*

The tool returned a research memo of approximately 800 words. The following excerpt contains a single cited authority and a single characterized holding — the target of this audit entry.

> Minnesota is an at-will employment state. An employer may terminate an at-will

> employee for any reason or no reason, provided the reason is not itself unlawful. *Andersen v. Northland Manufacturing, LLC*, 862 N.W.2d 231, 238 (Minn. Ct. App. 2021) (holding that an employer's decision to terminate an at-will employee is presumptively lawful and requires only a legitimate business reason). Exceptions include statutory anti-discrimination provisions under the Minnesota Human Rights Act, Minn. Stat. § 363A.08.

For the audit, three claims about *Andersen* are extracted:

1. **Existence claim.** *Andersen v. Northland Manufacturing, LLC* is reported at 862 N.W.2d 231, and page 238 says something related to the tool's proposition.
2. **Holding claim.** *Andersen* holds that at-will termination is "presumptively lawful and requires only a legitimate business reason."
3. **Validity claim.** *Andersen* remains good law on this point.

B.3 The Exists check

Retrieval.

Reviewer 1 opens the ground-truth corpus's citation resolver. The citation `862 N.W.2d 231` returns a single hit: *State v. Halverson*, 862 N.W.2d 231 (Minn. 2015). *State v. Halverson* is a criminal-procedure case about the admissibility of dashcam evidence. It is not the case the tool cited. No case by the name *Andersen v. Northland Manufacturing, LLC* is indexed at any reporter in the corpus.

Reviewer 2, blind to Reviewer 1, runs the same query and returns the same result.

Rubric application.

Both reviewers score the Exists check as **Fail — Fabricated** under Appendix A. The citation resolves to a real page of a real reporter, but not to the case the tool cited. The reporter and page number are decorative rather than referential.

Kappa contribution.

Agreement: 1. This authority contributes to a positive kappa numerator for the Exists check.

B.4 The Holds check

Because the Exists check failed, the Holds check is procedurally moot for this authority: there is no opinion to compare the tool's characterization against. Ratia records the Holds check as **N/A — pre-empted by Exists fail**. This is standard procedure and is documented in the scoring log to prevent later confusion.

Where an Exists failure is not fatal — for example, a citation where the case exists but the pincite is off by a page — the Holds check runs on the corrected pincite. That is not the situation here.

B.5 The Valid check

Same as the Holds check: the Valid check requires a real authority to check the validity of. Because the tool has not cited a real authority, there is nothing to validate. Recorded as **N/A — pre-empted by Exists fail**.

The audit report notes that the failure of the Exists check on this authority means no downstream check is possible. It does not mean the tool would have failed those checks if a real authority had been cited; it means the question could not be reached.

B.6 A counter-example on the Holds check

For the purpose of illustration, consider what the Holds check would look like if the tool had cited a real case incorrectly.

Suppose the tool had cited a real Minnesota Court of Appeals opinion — call it *[Real Case]*, [X] N.W.2d [Y], [Z] (Minn. Ct. App. [year]) — for the same "presumptively lawful; requires only a legitimate business reason" proposition, and suppose the real opinion actually says:

> "The at-will doctrine permits an employer to terminate an employee for good reason, bad reason, or no reason at all, subject to statutory and common-law exceptions. This does not mean the employer's decision is presumptively lawful; it means the employer's motive is not, by itself, a basis for a wrongful-termination claim absent one of the recognized exceptions."

Reviewer 1 would characterize the holding as: *At-will termination is not, on its own, a basis for wrongful-termination liability; motive alone is not actionable absent statutory or common-law exception.*

Reviewer 2, blind, would characterize it as: *An at-will employer is not liable for termination on the basis of motive alone; recognized exceptions must be pleaded.*

Both reviewers would compare the tool's characterization ("presumptively lawful and requires only a legitimate business reason") to their own.

- "Presumptively lawful" overstates the holding, which the opinion explicitly declined to frame that way.
- "Requires only a legitimate business reason" imports a burden-of-proof element the opinion does not adopt.

Rubric application. Both reviewers score this as **Partial — Overreach**: the tool's characterization is directionally correct but broader than the opinion supports, in a way that could mislead a reasonable attorney reader about the burden framework. Kappa agreement: 1.

Had the reviewers disagreed — one scoring Pass, one scoring Fail — a third senior reviewer would receive both scoring notes and the opinion and issue a written adjudication.

B.7 Excerpt from the scorecard

For the *Andersen v. Northland Manufacturing* entry, the scorecard for this authority reads:

Check	Score	Rubric level	Note
Exists	Fail	Fabricated	Cited reporter and page resolve to *State v. Halverson* (Minn. 2015); no such case as *Andersen v. Northland Mfg.* in the corpus.
Holds	N/A	Pre-empted	Cannot compare characterization to a non-existent opinion.
Valid	N/A	Pre-empted	Cannot validate a non-existent opinion.

B.8 Excerpt from the failure profile

The failure profile section of the Ratia Report groups this failure with other fabrications from the same audit period. A truncated entry:

> **Failure type: Fabricated citation.** Frequency in this audit: 4 of 47 audited citations (8.5%). Pattern: Court of Appeals-style citations in the 862-864 N.W.2d range, with plausible party-name pairings that do not resolve to real opinions. In every observed instance, the reporter and page number were real but attached to a different, unrelated case. This pattern is consistent with a retrieval failure or system-prompt weakness in which the tool constructs plausible-looking citations from citation-format templates rather than from a resolvable retrieval source.

The report does not speculate about the vendor's internal architecture. It reports the pattern, its frequency, and the type of downstream risk (a lawyer signing a filing containing this citation would be citing a case that does not exist).

B.9 Aggregate kappa

For the full audit period, the Cohen's kappa between Reviewer 1 and Reviewer 2, pre-adjudication:

Check	κ	Target	Notes
Exists	0.94	≥ 0.85	Nearly all disagreements were on Partial vs. Fail on citation-format ambiguities.
Holds	0.76	≥ 0.70	Disagreements were primarily on Overreach vs. Pass in overlapping-issue opinions. Rubric refinement scheduled for v1.1.
Valid	0.88	≥ 0.80	Highest agreement; the check is largely mechanical once the corpus is complete.

Adjudication rate: 11.5% of scored authorities required a third-reviewer adjudication.

B.10 What this example does not show

This example shows one authority's audit trace in enough detail to make the method concrete. A real Ratia Report contains dozens or hundreds of such traces, aggregated into the scorecard and failure profile described in §6. The example also does not illustrate:

- Retrieval-recall analysis (out of scope for grounding audits; separate methodology)
- Comparative analysis across multiple tools (a different report format)
- Vendor response and correction cycles (an operational, not methodological, feature)
- The full rubric text (Appendix A skeleton; full rubric published separately)

Appendix C — Related work and references

C.1 Reported sanctions cases and coverage

Ratia's methodology exists because the errors described in this paper are not hypothetical. The following reported matters are cited in the LinkedIn commentary that accompanies this paper and in §1.2 of the paper itself. Full pincites and case captions should be verified against the reporter of record before republication.

Foundational federal matter.

- *Mata v. Avianca, Inc.*, No. 22-cv-1461 (S.D.N.Y. June 22, 2023) (Castel, J.) (imposing sanctions on counsel for filing a brief containing citations to six nonexistent judicial decisions generated by ChatGPT). The seminal U.S. matter on AI-fabricated citations in legal filings.

State-court sanctions matters cited in this paper.

- Mississippi disqualification (2026). *See* Judge sanctions lawyers on both sides of a case over AI-fabricated citations, *N.Y. Times*, June 9, 2026. Underlying trial-court order to be added.
- New York appeals court $10,500 sanction (2026). *See* Attorney and law firm sanctioned $10,500 for AI-generated fake citations, *N.Y. Daily Record*, June 26, 2026. Underlying appellate order to be added.
- Sullivan & Cromwell forty-plus-error filing (2026). *See* Elite firm apologizes to judge for 40+ AI-fabricated errors in a court filing, *CNN Business*, April 23, 2026. Underlying filing and order to be added if the court's docket is available.

Aggregate empirical coverage.

- Courts have addressed AI errors in more than 1,400 cases — and lawyers keep citing fake law anyway, *Scientific American*, May 2026. The underlying dataset from which the "more than fourteen hundred filings" figure is drawn.

C.2 Editorial-service methodology

Ratia's Valid check overlaps with the function performed by commercial citator services. The following documentation describes those services' operation and, by comparison, clarifies where a pincite-grounded audit adds distinct information.

- Thomson Reuters, *KeyCite: How It Works* (product documentation; latest version). Describes the treatment flags (yellow, red, orange) and citing-references indexing used to flag negative treatment.
- LexisNexis, *Shepard's Signal Reference Guide* (product documentation; latest version). Analogous flagging system.

Where a citator flags an authority as "criticized" or "distinguished," the Valid check does not accept the flag as conclusive; the citing opinions are read to determine whether the authority remains authoritative on the specific point cited. See §3.3.

C.3 Legal-AI evaluation literature

Independent academic evaluation of legal-AI grounding is thin. The literature that exists tends to evaluate general legal reasoning (bar-exam performance, contract classification, statutory question-answering) rather than pincite-level grounding as defined in §2.1. Titles below are illustrative categories, not comprehensive; a full literature review is out of scope

for this methodology paper and is planned for a subsequent research note.

Bar-exam and general legal reasoning. Studies evaluating large language models on multistate bar examination questions and short-form legal reasoning tasks. These studies measure a different capacity than pincite grounding: they measure the model's ability to reason from stated premises to a legal conclusion, not the model's ability to accurately represent what a court held in a specific opinion at a specific pincite. High bar-exam performance is compatible with the failure modes described in §1.2. *[Specific citations to be added by author; see, e.g., work published by Stanford, Cornell, and University of Chicago legal-tech research groups.]*

Task-specific benchmarks. Legal benchmark suites evaluating classification, extraction, and question-answering tasks on legal documents. *[Specific citations to be added by author; e.g., LegalBench and its successors.]*

AI-hallucination in legal settings. Empirical studies documenting hallucination rates in legal-AI outputs, typically self-reported by product teams. Methodological limits of self-reported figures discussed in §1.1. *[Specific citations to be added by author.]*

C.4 The AI-model landscape

For readers unfamiliar with the "wrapper" framing in §1.3, the following primary references describe the general-purpose language models that underlie most current legal-AI products.

- Anthropic, *Claude* model family documentation (developer documentation; latest version). The Claude series is the model most frequently underlying legal-AI products marketed to U.S. firms as of the publication date.
- OpenAI, *GPT* model family documentation (developer documentation; latest version). The GPT-4 and successor models underlie a substantial share of legal-AI products.
- Google, *Gemini* model family documentation (developer documentation; latest version).

Nothing in this section should be read as endorsing any of these models for legal use. Each is a general-purpose model; grounding for legal use is a downstream engineering and evaluation problem that this paper is about.

C.5 Legal ethics and professional-responsibility guidance

Not surveyed in detail in this paper, but relevant to any firm considering how to communicate audit findings and adjust its practice. The following are entry points; each state bar has its own guidance and the following are U.S. federal or national.

- American Bar Association, *Formal Opinion 512* (July 29, 2024) (generative AI tools and the Model Rules of Professional Conduct). The ABA's first formal opinion on generative-AI use in law practice.
- State-bar advisory opinions on generative AI. *[Specific citations to be added by author.]* By publication date, more than two dozen state bars have issued formal or informal

guidance.

C.6 Suggested reading, methodology-adjacent

For readers interested in the epistemology underlying "grounding" as an audit target:

- Literature on evaluating machine-generated outputs against source documents (retrieval-augmented generation evaluation, citation-verification systems). *[Author to add specific references.]*
- Literature on inter-rater reliability in legal document review (relevant to §5.3). *[Author to add specific references.]*

Version 1.0. Author intends to expand §C.3, §C.5, and §C.6 in a subsequent revision as the underlying literature review is completed. Comments and suggestions on additions may be submitted via ratialegal.com.

Colophon

Ratia — from *ratio decidendi*, the reasoning that decides the case — provides independent audits of legal-AI grounding measured against the court-adjudicated record. Ratia is a measurement service; it is not the practice of law, does not provide legal advice, does not evaluate specific client matters, does not certify legal-AI products for general use, and forms no attorney-client relationship. Founder: Mikael Merissa (see ratialegal.com/about).

This paper is version 1.0, dated 2026-07-22. It supersedes no prior methodology. Comments and criticism are welcome at ratialegal.com.