PERSLIS RESEARCH

A Citation Floor for Case Law

The Mata v. Avianca *citations, re-checked against a free case database*

A model may propose a citation. Only a matching record in a real case database may admit it, and where the database cannot tell, the floor says so.

Perslis Research | September 2026

Research prototype · not legal advice · not a citator

Abstract. In *Mata v. Avianca, Inc.* (S.D.N.Y. 2023), a court sanctioned a filing that cited six decisions that do not exist. We describe a citation floor with one rule: a model may propose a case name and citation, but only a matching record in a public case-law database may admit it. The floor returns one of six verdicts (GROUNDED with a receipt, AMBIGUOUS, NOT_FOUND, MISCITED, NOT_GOOD_LAW, UNVERIFIABLE) and treats a miss as evidence of fabrication only in reporters it assumes the free database covers in full; elsewhere, and on every source error, it abstains. In one live run against CourtListener (eight queries, no API token), it admitted none of the five fabricated citations tested: it rejected three (two NOT_FOUND, one MISCITED) and held two Westlaw citations as UNVERIFIABLE, and it grounded three real Supreme Court citations with source links. Its MISCITED receipt names the real case the court found at that citation. The run covered 5 of the 12 fabricated or not-as-cited citations in the opinion, and none of the 6 misused real decisions, which an existence check admits by design. Reading the code, we find three failure paths the run did not exercise: a fabricated caption sharing a common party word with the real case would be admitted; a real case written in a variant form can be rejected; and the good-law veto is not on the path tested. We state the coverage assumption that makes rejection sound, and design the next experiment with its gate fixed in advance.

1 Introduction

A case citation is the part of a legal filing that can be checked mechanically. It names a reporter, a volume and a first page; either an opinion begins there or none does, and if one does, it has a name. Generative models write text that looks like citations whether or not the opinions exist, and the legal record now shows what follows. In *Mata v. Avianca, Inc.* the court found six decisions cited in an opposition to a motion to dismiss to be fake, and its findings record the acknowledgment that they "were generated by ChatGPT and do not exist" [1, ¶ 36]. It found that one of the fake opinions "includes internal citations and quotes from decisions that are themselves non-existent" [1, ¶ 29], and it ordered sanctions. The record also contains the failure in its simplest form. Asked whether the cases were real, the tool "responded that it had supplied 'real' authorities that could be found through Westlaw, LexisNexis and the Federal Reporter" [1, ¶ 45]. A proposer's assurance about its own proposal is not a check.

This paper describes a check that sits outside the proposer. We call it a *citation floor*, after the admission pattern of our earlier work: a model may propose, and only a grounded source may admit [28, 26, 22]. The floor asks two questions of each citation a model proposes. Is there a real opinion at this citation? Is it the case the proposer names? It answers from a public case-law database, CourtListener [9], in one of six verdicts. Each verdict carries an action (admit, reject or hold) and, where a record exists, a *receipt*: the record itself, with its link. The design rests on an asymmetry that lawyers already know. Calling real law fabricated is a false accusation, and it can do as much harm as admitting a fake. So a lookup that finds nothing counts as evidence of fabrication only where the database's coverage makes absence informative.

Everywhere else, and on every error from the source, the floor abstains.

We test it where the problem became public. We ran the floor once, live, on the fabricated citations from *Mata* that its demonstration script contains, and on three real Supreme Court citations as controls. We then set the result against the court's own findings, citation by citation, and against the full list of authorities the opinion identifies, most of which the script does not contain. Finally we read the code for ways the contract could fail on inputs the run did not include.

1.1 Contributions

C1. A contract for admitting citations (§3). Six verdicts, each tied to an action and a receipt, and a coverage rule under which a database miss counts as evidence of fabrication only for reporters the floor treats as fully covered. We state what a rejection certifies (Proposition 1), and we show that a transport failure can never become a verdict of existence or non-existence (Proposition 2).

C2. A re-check on the public record (§4–§5). In one live run (27 September 2026, eight queries, no API token) the floor admitted **0 of 5** fabricated citations: it rejected **3** (two NOT_FOUND, one MISCITED) and held **2** as UNVERIFIABLE. It grounded **3 of 3** real citations with receipts. The one MISCITED receipt names the real case the court found at that citation, *United States v. ISS Marine Services* [1, ¶ 34].

C3. A map of the whole record (§6). The opinion identifies 12 fabricated or not-as-cited citations and 6 real decisions cited for what they do not say. We sort all 18 into six error classes and state, for each class, what the contract can say, what was run (**5 of 18**), and what no existence check can catch.

C4. An audit and a next experiment (§7–§8). Reading the code, we find three failure paths the run did not reach: a false admission through a shared party word, false rejections of real law written in variant forms, and a good-law veto that is absent from the path tested. From these we design a failure-directed comparison of deterministic and learned verification with its gate fixed in advance.

1.2 What we claim, and what we do not

> On eight inputs, on one day, against one free database, the floor admitted no fabricated citation and rejected no real one. Its receipts can be checked against the court's own findings, and one of them matches a finding exactly.

We do not claim a false-admission rate or a false-rejection rate: the sample is five fabricated citations and three famous real ones, and the audit in §7 shows inputs on which the floor would fail. A GROUNDED verdict means that the citation exists and names the claimed case. It does not mean that the case is still good law, and it does not mean that the case supports the proposition it is cited for. The floor is not a citator, and nothing here is legal advice. We did not evaluate any commercial research tool. None of the ingredients is new: citation lookup services, citation parsers and citators exist (§9), and CourtListener itself offers a token-based lookup described as "a guardrail to help prevent hallucinated citations" [8]. Our contribution is the contract around the lookup: what a miss is allowed to mean, what happens on error, what a verdict carries, and an honest measurement of how far the contract reaches on the one public record where the problem is best documented.

2 The public record

We rely only on the court's Opinion and Order on Sanctions, *Mata v. Avianca, Inc.*, No. 22-cv-1461 (PKC), ECF No. 54 (S.D.N.Y. June 22, 2023), a 43-page filing: 34 pages of opinion and two appendices [1]. We obtained it from the public RECAP archive and pinned its hash (Appendix B). Paragraph numbers below are the opinion's findings of fact. We name the parties and the court as they appear in the case citation, and nobody else.

What happened. The plaintiff sued an airline over an injury on an international flight. The airline moved to dismiss the claims as time-barred under the Montreal Convention (¶ 3). The affirmation opposing that motion, filed on 1 March 2023, cited decisions the airline's counsel could not find: its reply of 15 March "detailed by name and citation seven purported 'decisions' " it could not locate (¶ 7). On 11 and 12 April the court ordered copies of nine cited decisions (¶¶ 13–14). Purported copies or excerpts of all but one were filed on 25 April (¶ 18). The court then made findings on the "decisions" (¶¶ 24–36) and on 22 June 2023 ordered sanctions under Rule 11 of the Federal Rules of Civil Procedure or, alternatively, its inherent authority (at 34).

What the court found. Table 1 lists every authority the opinion identifies, in three groups.

- *Six fabricated decisions* cited in the affirmation, which the findings record "do not exist" (¶ 36). For three of them the court states where their citations actually lead. The Federal Supplement citation given for *Petersen* belongs to a different, real case (¶ 34). The Federal Reporter citations given for *Varghese* and *Miller* are, in the court's words, "associated with" and "to" other real opinions, which begin on earlier pages of the same volumes (¶¶ 28, 32). The other three (*Shaboon*, *Martinez*, *Durden*) "contain similar deficiencies" (¶ 35); they are cited by a public-domain Illinois citation and two Westlaw numbers.
- *Six authorities cited inside the fake Varghese opinion that do not exist, or not as cited* (¶ 29 a–f). For each, the court states what is actually at the cited citation. Two of them borrow the names of real cases decided under different citations: *Zicherman* (a real Supreme Court decision) and *In re BDC 56 LLC* (a real Second Circuit decision). *Zicherman*, with its false Federal Reporter citation, was also cited in the affirmation itself (¶ 14).
- *Six real decisions*, correctly named and cited, that "do not contain the language quoted or support the propositions for which they are offered" (¶ 29 g). One is also given the wrong court.

The record also shows why a check must sit outside the proposer. When the tool was asked whether the cases were real, it answered that they were and could be found in Westlaw, LexisNexis and the Federal Reporter (¶ 45). It also reflects testimony that free sites exist where a known citation to a reported decision can be entered and the decision displayed (¶ 12). The check this paper automates is one anyone can run by hand. The floor's purpose is to make that check mandatory, and to make its answer binding and receipted.

Table 1. Every authority the court's opinion identifies as fabricated, not as cited, or misused, with the court's finding, the error class we assign (Section 6) and the floor's verdict where it was run. Citations are as the opinion gives them. Each row's exact quotation and line in the opinion text are in `analysis/mata-record.json`.

#	Authority, as cited	The court's finding	Opinion	Class	Floor (27 Sep.)
Fabricated decisions cited in the affirmation (¶ 36)					
1	*Varghese v. China Southern Airlines Co., Ltd.*, 925 F.3d 1339 (11th Cir. 2019)	No such decision (the Eleventh Circuit clerk). The citation "is associated with" *J.D. v. Azar*, 925 F.3d 1291 (D.C. Cir. 2019).	¶¶ 26–28	A	NOT_FOUND
2	*Shaboon v. Egyptair*, 2013 IL App (1st) 111279-U (Ill. App. Ct. 2013)	Does not exist; "similar deficiencies".	¶¶ 35–36	C	not run
3	*Peterson v. Iran Air*, 905 F. Supp. 2d 121 (D.D.C. 2012)	Does not exist. The citation is to *United States v. ISS Marine Services*, 905 F. Supp. 2d 121 (D.D.C. 2012). Listed as "Peterson" in the order, annexed as "Petersen".	¶¶ 33–34	B	MISCITED
4	*Martinez v. Delta Airlines, Inc.*, 2019 WL 4639462 (Tex. App. Sept. 25, 2019)	Does not exist; "similar deficiencies".	¶¶ 35–36	C	UNVERIFIABLE
5	*Estate of Durden v. KLM Royal Dutch Airlines*, 2017 WL 2418825 (Ga. Ct. App. June 5, 2017)	Does not exist; "similar deficiencies".	¶¶ 35–36	C	UNVERIFIABLE
6	*Miller v. United Airlines, Inc.*, 174 F.3d 366, 371-72 (2d Cir. 1999)	Does not exist. The citation is to *Greenleaf v. Garlock, Inc.*, 174 F.3d 352 (3d Cir. 1999).	¶¶ 30–32	A	NOT_FOUND
Cited inside the fake Varghese*: do not exist, or not as cited (¶ 29 a–f)*					
7	*Holliday v. Atl. Capital Corp.*, 738 F.2d 1153 (11th Cir. 1984)	Does not exist. The case at that citation is *Gibbs v. Maxwell House*, 738 F.2d 1153 (11th Cir. 1984).	¶ 29(a)	B	not run
8	*Gen. Wire Spring Co. v. O'Neal Steel, Inc.*, 556 F.2d 713, 716 (5th Cir. 1977)	Does not exist. The case at that citation is *United States v. Clerkley*, 556 F.2d 709 (4th Cir. 1977).	¶ 29(b)	A	not run
9	*Hyatt v. N. Cent. Airlines*, 92 F.3d 1074 (11th Cir. 1996)	Does not exist. Two brief orders in other cases appear at 92 F.3d 1074.	¶ 29(c)	B	not run
10	*Zaunbrecher v. Transocean Offshore Deepwater Drilling, Inc.*, 772 F.3d 1278, 1283 (11th Cir. 2014)	Does not exist. The case at that citation is *Witt v. Metropolitan Life Ins. Co.*, 772 F.3d 1269 (11th Cir. 2014).	¶ 29(d)	A	not run
11	*Zicherman v. Korean Air Lines Co.*, 516 F.3d 1237, 1254 (11th Cir. 2008)	Not as cited: the real *Zicherman* is 516 U.S. 217 (1996); the F.3d citation is *Miccosukee Tribe v. United States*, 516 F.3d 1235. Also cited in the affirmation (¶ 14).	¶ 29(e); also ¶ 14	A+D	not run
12	*In re BDC 56 LLC*, 330 B.R. 466, 471 (Bankr. D.N.H. 2005)	Not as cited: a Second Circuit decision of that name is 330 F.3d 111 (2d Cir. 2003); the case at the B.R. citation is *In re 652 West 160th LLC*, 330 B.R. 455.	¶ 29(f)	A+D	not run
Cited inside the fake Varghese*: real, but not what they are cited for (¶ 29 g)*					
13	*In re Rimstat, Ltd.*, 212 F.3d 1039 (7th Cir. 2000)	Real; concerns Rule 11 sanctions and does not discuss the bankruptcy stay.	¶ 29(g)	E	not run
14	*In re PPI Enterprises (U.S.), Inc.*, 324 F.3d 197 (3d Cir. 2003)	Real; does not discuss the stay; misidentified as a Second Circuit opinion.	¶ 29(g)	E+F	not run
15	*Begier v. I.R.S.*, 496 U.S. 53 (1990)	Real; does not discuss the stay.	¶ 29(g)	E	not run
16	*Kaiser Steel Corp. v. W. S. Ranch Co.*, 391 U.S. 593 (1968)	Real; does not discuss the stay.	¶ 29(g)	E	not run
17	*El Al Israel Airlines, Ltd. v. Tsui Yuan Tseng*, 525 U.S. 155 (1999)	Real; does not contain the quoted language.	¶ 29(g)	E	not run
18	*In re Gandy*, 299 F.3d 489 (5th Cir. 2002)	Real; affirmed a denial of a motion to compel arbitration.	¶ 29(g)	E	not run

3 The floor's contract

We describe the floor by what it accepts, what it answers and what each answer carries, not by its internals. Figure 1 draws the path the *Mata* test uses.

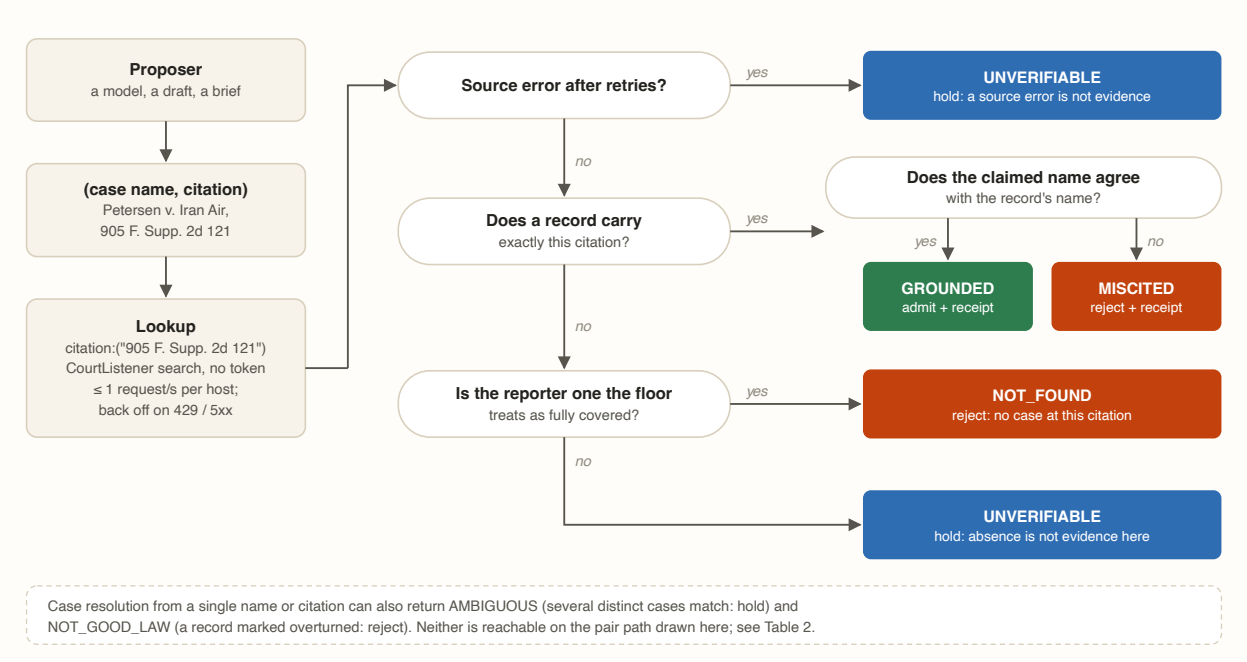

Figure 1. The pair check, `verify(name, citation)`. A proposer supplies a case name and a citation; the floor issues one exact query to a public case-law database and returns a verdict with an action. Green admits, orange rejects, blue holds for a person or a paid citator. A record-level receipt accompanies GROUNDED and MISCITED. The two verdicts in the dashed box come only from resolving a single reference (Table 2).

3.1 Inputs and the database

The floor has two entry points. `resolve(reference)` takes one case name *or* one citation. `verify(name, citation)` takes the pair a brief actually writes, *Petersen v. Iran Air, 905 F. Supp. 2d 121*, and asks whether the citation exists and belongs to the named case. A citation must be given in its bare form: volume, reporter, first page. Pin cites and court-and-year parentheticals are not stripped by the floor, and §7 shows what happens when they are left in.

The database is CourtListener, run by the Free Law Project [9]. The floor uses only its public search interface, with no account and no token. CourtListener documents `citation` as "an all-encompassing field containing all of the citations for an opinion", and documents quotation marks as a phrase search that "will not apply stemming or match synonyms" [7]. The floor therefore asks for one quoted phrase inside that field, `citation:("905 F. Supp. 2d 121")`, rather than a free-text search. The lane's development notes record that an unquoted field query matched volume and page separately; we did not re-test that here. The database's ranking never decides a verdict by itself: of the records returned, the floor keeps only those that list the input among their citations, compared after case and spacing are normalised.

3.2 Verdicts, actions and receipts

Table 2 is the whole output vocabulary. Each verdict maps to one of three actions: *admit* (the citation may enter trusted state), *reject* (it may not), or *hold* (it may not yet, and a person or a paid citator must decide). A *receipt* is the database record behind a verdict: the case name, every citation the database lists for it, the

court, the date, the record's identifier and its URL. Anyone can open a receipt and check it. Receipts are links to a public record, not signed evidence, and the record behind a link can change (§10).

Table 2. The floor's verdicts. "Pair" is `verify(name, citation)`; "single" is `resolve(reference)`. The last column records whether the verdict occurs in our evidence: the live run of §5, or the lane's committed tests, whose last recorded run (23 September) left no failures; we did not re-run them.

Verdict	Meaning	Action	Entry	In our evidence
GROUNDED	A real record carries the citation and the claimed name agrees with it (pair); or exactly one record matches the reference (single).	admit, with receipt	pair, single	run: 3; tests
AMBIGUOUS	Several distinct records match a single reference.	hold: disambiguate	single	run: 0; tests allow it
NOT_FOUND	No record carries the citation, and its reporter is one the floor treats as fully covered; or no record matches a name.	reject	pair, single	run: 2; tests
MISCITED	A real record carries the citation, but it names a different case.	reject, with receipt	pair	run: 1; tests
NOT_GOOD_LAW	Every matching record is marked as negatively treated (overruled, vacated).	reject (hard veto)	single	run: 0; no test
UNVERIFIABLE	The source failed after retries; or no record carries a citation in a form the free database does not fully cover (Westlaw and Lexis numbers, neutral citations, other reporters).	hold	pair, single	run: 2; tests

3.3 When is absence evidence?

The central design choice is what a miss is allowed to mean. A lookup that finds nothing proves nothing on its own. The citation may be fabricated, the database may lack the opinion, or the request may have failed. The floor reads a miss as fabrication only when the first two can be separated, and that depends on coverage.

Definition 1 (Covered reporters)**.** Let D be the database and R_c the set of reporters the floor treats as fully covered: the United States Reports, the Supreme Court Reporter, the Lawyers' Edition, the Federal Reporter series and the Federal Supplement series. A canonical citation $c = (v, r, p)$ names volume v of reporter r and first page p. *Coverage* is the assumption that for every $r \in R_c$, every opinion published in r that begins at page p of volume v has a record in D listing (v, r, p) among its citations. *Retrieval* is the assumption that the quoted query for c returns every such record among the results the floor examines.

Proposition 1 (What a rejection certifies)**.** *Under coverage and retrieval, if the pair check returns* NOT_FOUND *for a canonical citation $c = (v, r, p)$ with $r \in R_c$, then no opinion published in r begins at page p of volume v.*

Proof. If such an opinion existed, coverage would give a record listing c, and retrieval would place it among the results examined. The floor would then find a record carrying c and return GROUNDED or MISCITED, not NOT_FOUND. ∎

The proposition is short on purpose. It makes the assumptions visible. A rejection is exactly as sound as three things: the coverage claim for R_c, the retrieval claim for the query, and the input being canonical.

None of the three is measured in this paper; §7 and §10 say how each can fail. The floor's own note on a rejection says "fabricated". Strictly, what it certifies is that no opinion begins at the cited page. Fabrication is the inference the court drew from the whole record.

Proposition 2 (A transport failure is never a verdict)**.** *If the database cannot be reached, answers with an HTTP error status, or returns a body that is not JSON, and the failure persists through the retries of §3.4, the verdict is* UNVERIFIABLE*. No such failure can produce* NOT_FOUND *or* GROUNDED*.*

Argument. By construction. Every request goes through one function that either returns parsed JSON or raises a single error type after its retries. Both entry points catch that error type and return UNVERIFIABLE before any verdict logic runs. A failure of another kind, such as a response cut off mid-read, raises an exception and yields no verdict at all. We checked this by reading the code (Appendix B gives the lines), not by injecting faults. ∎

The proposition does not cover a well-formed response that is wrong. An empty result set returned in error, or a renamed field in the database's answers, would turn real citations into NOT_FOUND. The guard against that is to run known real citations alongside every batch, as canaries, and to void the batch if any of them fails to ground. The three controls in our run play that role (§4).

Remark 1 (Why hold rather than guess)**.** The floor faces a three-way decision with a reject option [5]. Admitting a fabrication and rejecting real law are both costly errors; holding costs a person's time. A miss in a covered reporter is decisive under the assumptions of Proposition 1, so rejecting is right. A miss on a Westlaw number, whose coverage the free database does not claim, says little either way, so holding is right. The contract spends abstentions to avoid both kinds of error. It does not trade one for the other.

3.4 Throttling and fail-fast

The lane's development notes record that CourtListener throttles unauthenticated search heavily. Whatever the rate, a rate limit is not evidence about the law. The floor spaces its requests at least one second apart per host. On HTTP 429, 502, 503 or 504, and on network errors, it retries up to four attempts in all. It waits for the server's `Retry-After` value if one is given, otherwise 1, 2 and 4 seconds, and it never waits more than 30 seconds at a time. Any other HTTP error fails at once. So does a body that is not JSON. A failure that survives the retries becomes UNVERIFIABLE (Proposition 2). The floor never falls back to a guess, a cached answer or another source.

3.5 Good law is not checked yet

A citation can exist, name the right case and still not be good law. Knowing that a case was overruled, vacated or superseded requires the treatment history that commercial citators such as Shepard's and KeyCite provide; free search results do not expose it. The floor has the veto (NOT_GOOD_LAW), but its free-tier signal is limited to what a record itself shows, and the pair check used in the *Mata* test does not consult it (§7). Neither our run nor the committed tests produced this verdict. GROUNDED therefore means "exists, and names the claimed case", never "good law". Closing this gap needs either a paid citator or treatment built from CourtListener's free citation network. The second is planned and not built.

3.6 Implementation and scope

The floor is 344 lines of Python across three modules and uses only the standard library. A 123-line live test file and an 86-line demonstration script accompany it, all at commit `533f783`. The Admissible Motion paper names the legal floor as one instance of a cross-domain admission pattern, alongside a bio floor and a motion floor [22]. The pattern is shared: propose, check against a grounded source, admit, reject or hold, with receipts. The code is not. The legal floor imports nothing from the bio floor's engine; it is

a separate implementation of the same contract. The connector also implements CourtListener's token-gated citation-lookup endpoint, but the floor never calls it. Everything here is United States case law, and every reporter in the covered set is a federal or Supreme Court reporter.

4 Method

One run, recorded. We ran the lane's demonstration script once, live, on 27 September 2026 from 09:08:25 to 09:08:33 UTC. No CourtListener token was set. The code was at commit `533f783`, and the script's directory was byte-identical to that commit (no local changes). We ran it with Python 3.9.6, with bytecode writing disabled so that the run left nothing in the lane's directory. Its full output, with a SHA-256 hash, and the hashes of every source file are in the paper's `evidence/` folder (Appendix B). The script makes one database query per input, eight in all. We did not repeat the run. The paper's scope allowed one live run, both to keep the evidence fixed and to keep the load on a free public service small.

Inputs. The script passes eight (name, citation) pairs to `verify`.

- *Five fabricated pairs from the record*: *Varghese*, *Petersen*, *Miller*, *Martinez* and *Durden*. They are five of the six fabricated decisions in Table 1. *Shaboon* is not in the script. The citations were normalised by hand to the bare form the floor requires. *Miller*'s pin cite ("371–72") and every court-and-year parenthetical were dropped. *Petersen* is spelled as in the annexed text; the court's order lists it as "Peterson".
- *Three real controls*: *Miranda v. Arizona*, 384 U.S. 436; *Brown v. Board of Education*, 347 U.S. 483; and *Gideon v. Wainwright*, 372 U.S. 335. They serve as canaries (§3.3). They are also the easiest possible real citations: landmark decisions in the United States Reports.

Outcome coding. We code `GROUNDED` as *admitted*, `NOT_FOUND` and `MISCITED` as *rejected*, and `UNVERIFIABLE` as *held*. For a fabricated input, admission is the failure the floor exists to prevent. A rejection is a catch. A hold is neither: it is a hand-off, and we never count it as a catch. For a real input, rejection is a false accusation and a hold is a cost. A small script re-parses the run's output, recomputes every count, and checks them against the demonstration script's own score lines. It aborts on any disagreement.

External check. For every verdict we compare the receipt, or its absence, with what the court found at the same citation (Table 1). A second script pins each authority, and each finding we rely on, to its line in the text of the opinion, and it aborts if any quoted fragment is missing. The error classes in §6 are ours. We assign them from the court's findings, not from the floor's output.

Committed tests, not re-run. The lane has 11 live tests. Its test cache shows that the last recorded session, at 18:38 EDT on 23 September, collected all 11 and recorded no failures. Tests that skip for lack of network would also leave the failure list empty, so the cache alone does not show they passed. The commit message reports them passing live. We cite them only as the lane's stated expectations. Four of them assert the *Mata* verdicts, and a fifth asserts the *Miranda* pair. Our run reproduces all five. The other six exercise single-reference resolution (five tests) and the citation-format detector (one), which the run did not use.

Audit by reading. We read every line of the floor and its connector and asked, for each verdict, what the verdict certifies and what inputs could make it wrong. Each finding in §7 cites the file and lines that support it. None was executed. The scope of this paper permitted the one live run and nothing else, so each audit finding is stated as a property of the code, with the input that would expose it.

5 Results

5.1 The run

Table 3 gives every verdict with its receipt or note. The demonstration script's own score lines agree with our recount.

> **Fabricated (n = 5): 0** admitted; **3** rejected (NOT_FOUND 2, MISCITED 1); **2** held (UNVERIFIABLE).
> **Real (n = 3): 3** admitted (GROUNDED), each with a CourtListener receipt; **0** rejected; **0** held.
> Eight queries in 8 seconds of wall-clock time, one run, no API token.

Two of the five fabricated citations were held, not caught. The floor's contract counts that as correct behaviour, because the free database does not claim to cover Westlaw numbers, but a hold catches nothing. It hands the citation to a person, who in this case would have needed Westlaw or the issuing courts to settle the question. The demonstration script's docstring says the floor "catches every one". Its score line reports "3/5 (+2 UNVERIFIABLE)", and the score line is the accurate one.

5.2 Agreement with the court

For each input, the floor's answer can be set against the court's finding at the same citation.

- ***Petersen*, 905 F. Supp. 2d 121: exact agreement.** The floor returned MISCITED with a receipt naming *United States of America v. Iss Marine Services, Inc.* (CourtListener's capitalisation) at that citation. The court found: "The Federal Supplement citation is to United States v. ISS Marine Services, 905 F. Supp. 2d 121 (D.D.C. 2012)" [1, ¶ 34]. The same case, found independently by a lookup that knew nothing of the opinion.
- ***Varghese*, 925 F.3d 1339, and *Miller*, 174 F.3d 366: the same rejection, a narrower reason.** The floor reported that no record carries either citation. The court associated them with real opinions that begin earlier in the same volumes, *J.D. v. Azar* at 925 F.3d 1291 and *Greenleaf v. Garlock* at 174 F.3d 352 (¶¶ 28, 32). The two accounts are consistent: a citation identifies an opinion by its first page, and no opinion begins at either cited page. The floor cannot say which opinion a cited page falls inside, because it resolves first pages only. A page-range lookup could (§8).
- ***Martinez* and *Durden* (Westlaw numbers): held, and rightly not rejected.** The court found both non-existent (¶¶ 35–36). The free database holds no record with either Westlaw number, and the floor abstained. The abstention is correct in a way that is easy to overlook. CourtListener does record some Westlaw numbers as parallel citations. The *ISS Marine* receipt above lists 2012 WL 5873682. So Westlaw numbers are partly indexed, and a miss on one is a coverage gap, not evidence that the case does not exist.

5.3 How much of the record was checked

Figure 2 places the run against the whole record. The run covered **5 of the 6** fabricated decisions. The opinion identifies **12** fabricated or not-as-cited citations in all (the six decisions and the six inside the fake *Varghese*), so the run covered **5 of 12**. It covered **0 of the 6** real decisions the opinion found misused. Across all **18** authorities the opinion identifies, **5** were run and **13** were not. The accurate one-line summary is therefore: *of the case's fabricated citations, the floor classified five, rejecting three and holding two, and admitted none.* It is not "the floor catches the *Mata* fabrications".

5.4 Receipts

All four verdicts backed by a record, the three GROUNDED and the one MISCITED, carried a receipt: the case name, every citation the database lists for the record, and a CourtListener URL. The receipts differ in how complete they are. *Miranda*'s record lists 7 citations and *Brown*'s 6, including their Supreme Court

Table 3. The single live run (27 September 2026, 09:08:25–09:08:33 UTC, no API token, code at commit `533f783`). Inputs are the name and bare citation the demonstration script passes; the receipt is the CourtListener record the floor returned. The last column is the court's finding (Table 1) or, for the controls, blank.

Input (name, citation)	Verdict	Receipt or note returned by the floor	The court's finding
Varghese v. China Southern Airlines Co., Ltd., 925 F.3d 1339	NOT_FOUND	note (verbatim): "no case carries this mainstream-reporter citation — fabricated, REJECTED"	No such decision (the Eleventh Circuit clerk). The citation "is associated with" *J.D. v. Azar*, 925 F.3d 1291 (D.C. Cir. 2019).
Petersen v. Iran Air, 905 F. Supp. 2d 121	MISCITED	real citation, different case: *United States of America v. Iss Marine Services, Inc.* — 905 F. Supp. 2d 121; 84 Fed. R. Serv. 3d 384; 2012 WL 5873682; 2012 U.S. Dist. LEXIS 166088 — https://www.courtlistener.com/opinion/2661490/united-states-of-america-v-iss-marine-services-inc/	Does not exist. The citation is to *United States v. ISS Marine Services*, 905 F. Supp. 2d 121 (D.D.C. 2012). Listed as "Peterson" in the order, annexed as "Petersen".
Miller v. United Airlines, Inc., 174 F.3d 366	NOT_FOUND	note (verbatim): "no case carries this mainstream-reporter citation — fabricated, REJECTED"	Does not exist. The citation is to *Greenleaf v. Garlock, Inc.*, 174 F.3d 352 (3d Cir. 1999).
Martinez v. Delta Airlines, Inc., 2019 WL 4639462	UNVERIFIABLE	note (verbatim): "citation not in the free corpus (e.g. Westlaw/unpublished) — cannot verify"	Does not exist; "similar deficiencies".
Estate of Durden v. KLM Royal Dutch Airlines, 2017 WL 2418825	UNVERIFIABLE	note (verbatim): "citation not in the free corpus (e.g. Westlaw/unpublished) — cannot verify"	Does not exist; "similar deficiencies".
Miranda v. Arizona, 384 U.S. 436	GROUNDED	*Miranda v. Arizona* — 16 L. Ed. 2d 694; 86 S. Ct. 1602; 384 U.S. 436; 1966 U.S. LEXIS 2817; 10 Ohio Misc. 9; 36 Ohio Op. 2d 237; 10 A.L.R. 3d 974 — https://www.courtlistener.com/opinion/107252/miranda-v-arizona/	(real control)
Brown v. Board of Education, 347 U.S. 483	GROUNDED	*Brown v. Board of Education* — 347 U.S. 483; 74 S. Ct. 686; 1954 U.S. LEXIS 2094; 38 A.L.R. 2d 1180; 53 Ohio Op. 326; 98 L. Ed. 873 — https://www.courtlistener.com/opinion/105221/brown-v-board-of-education/	(real control)
Gideon v. Wainwright, 372 U.S. 335	GROUNDED	*Gideon v. Wainwright* — 372 U.S. 335 — https://www.courtlistener.com/opinion/8954562/gideon-v-wainwright/	(real control)

Reporter and Lawyers' Edition parallels. *Gideon*'s lists only 372 U.S. 335, so a brief citing *Gideon* by a parallel citation would not have matched this record. Whether another record carries it, we did not check. The two NOT_FOUND and two UNVERIFIABLE verdicts carry only the floor's note, because there is no record to show.

5.5 Cost

The run took 8 seconds of wall-clock time for 8 queries. That is consistent with the floor's one-request-per-second spacing. We did not instrument retries or per-query latency, so we report no latency figure. The run needed no account and no token.

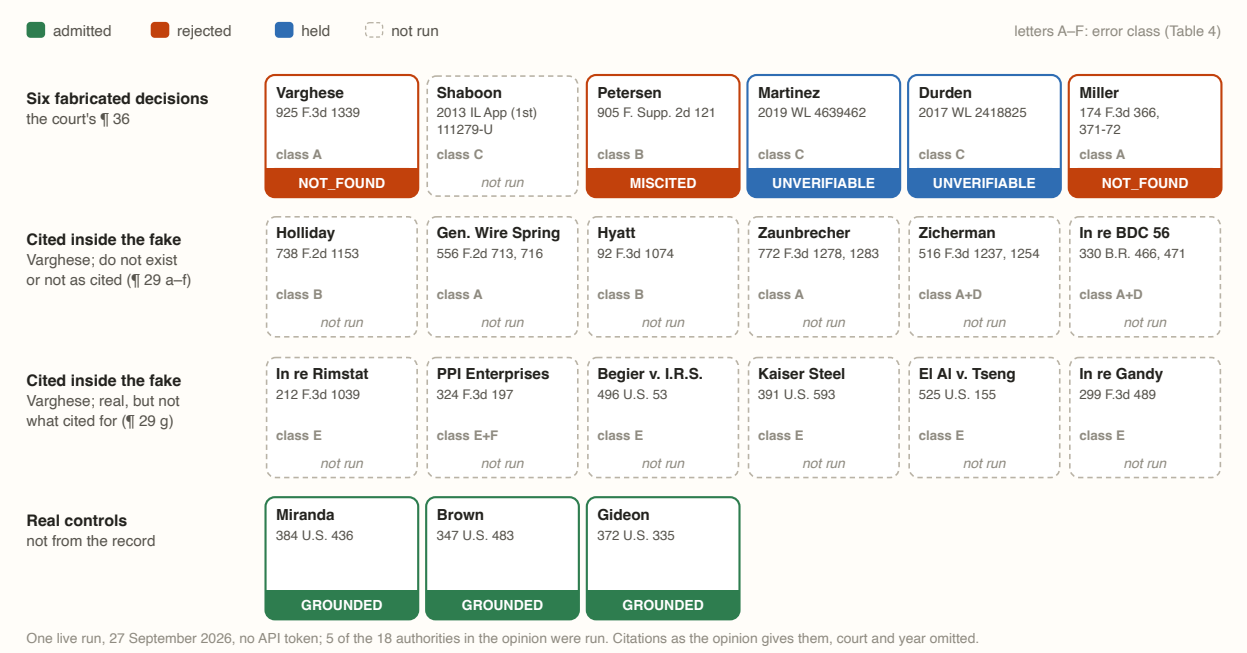

Figure 2. Every authority in the court's opinion, and what was actually run. Filled cells were run on 27 September 2026; dashed cells were not. Colour is the floor's action (admit, reject, hold), and every cell also prints its verdict. Letters are the error classes of Table 4. The three real controls are not from the record.

6 What an existence check can and cannot see

The eighteen authorities in the opinion fail in different ways. Table 4 sorts them into six classes, taken from the court's findings, and gives the verdict the contract produces for each class. Where a class was run, the verdict is measured. Where it was not, the verdict is what the contract of §3 prescribes, and we label it so: a prescription is not a result.

Classes A and B are what a citation floor is for. They are errors of identity, and a lookup settles them: either an opinion begins at the citation or none does, and if one does, its name either agrees with the claim or it does not. The three measured rejections fall here. So do six of the seven fabricated citations that were not run, five of them in covered reporters, which is why they are the first additions to the test set (§8).

Class C is where honesty costs coverage. A floor that rejected every Westlaw miss would have "caught" *Martinez* and *Durden*. The same rule would also accuse every real unpublished decision absent from the free corpus of being fabricated. The contract refuses that trade, and the price is two holds out of five. Paying for a citator removes the price. A better free rule does not.

Class D is why the unit of verification is the pair. *Zicherman v. Korean Air Lines Co.* is a real Supreme Court case [1, ¶ 29(e)]. A check that verified names would admit the fabricated Federal Reporter citation attached to it. The floor's pair check asks whether an opinion begins at the cited page, whatever the name.

Class E is out of reach for any existence check, including ours. Six of the eighteen authorities are real, correctly cited decisions offered for things they do not say. The floor would ground all six, correctly by its contract and uselessly for the filing. Magesh et al. call such a citation *misgrounded*: a response is misgrounded when its propositions "are cited but misinterpret the source or reference an inapplicable

Table 4. Error classes in the *Mata* record, what the floor's contract says about each, and what was measured. Counts are of the 18 authorities in Table 1; two authorities belong to two classes (A and D), and one to E and F. "Prescribed" is the verdict the contract specifies for that input, not an observed outcome.

	Class (as the court describes it)	Authorities	Floor	What would be needed
A	No opinion begins at the cited page; the court finds an opinion that begins earlier.	6: *Varghese*, *Miller*, *Gen. Wire Spring*, *Zaunbrecher*, *Zicherman*, *In re BDC 56*	NOT_FOUND in a covered reporter (measured 2 of 2 run); UNVERIFIABLE for *BDC 56*, whose reporter (B.R.) is not covered (prescribed).	A page-range lookup, to name the opinion the page falls in, as the court did.
B	The cited first page begins a different, real opinion.	3: *Petersen*, *Holliday*, *Hyatt*	MISCITED (measured 1 of 1 run). For *Hyatt*, two orders share the page, and the floor compares the name with the first record only (§7).	Compare the name with every record at the page.
C	A Westlaw number or a public-domain neutral citation.	3: *Martinez*, *Durden*, *Shaboon*	UNVERIFIABLE (measured 2 of 2 run); *Shaboon*, not run, depends on whether the free database indexes that neutral citation.	A paid citator, or the issuing court's own records.
D	A real case's name, attached to a citation it does not have.	2: *Zicherman*, *In re BDC 56*	Decided by the citation, as in class A (prescribed). A lookup by *name alone* would find the real case and admit it.	Always verify the pair, never the name alone.
E	A real decision, correctly cited, that does not contain the quotation or support the proposition.	6: ¶ 29(g)	GROUNDED, by design (prescribed). Existence is not support.	Read the opinion against the proposition: a support check.
F	The wrong court in the parenthetical.	1: *In re PPI Enterprises*	Not examined: the input carries no court.	Compare court and year with the record.

source" [16]. In their sense our GROUNDED is weaker than "grounded". It asserts identity, not support. A floor for support would have to read the opinion. That is a different check, with a different failure profile, and nothing in this paper measures it.

7 Audit: where the contract can fail

The run shows the floor behaving correctly on eight inputs. It does not show that the contract holds on inputs the run did not contain. We read the code for such inputs. Every finding below is a property of the code at commit `533f783`, with the file and lines that establish it. None was executed: this paper's scope allowed one live run and nothing else. Each is stated with the input that would expose it, and each input is in the test set of §8.

7.1 A false admission through a shared party word

The pair check compares the claimed name with the record's name leniently, by design. Its job is to catch the gross mismatch, a real citation with a different case, and not to fail on formatting (`floor.py`, lines 129–138). In practice it treats two names as agreeing when they share one distinctive party word (lines 29–30 and 162–164). Two inputs the run did not contain show what that admits:

- *Smith v. Board of Education*, 347 U.S. 483. The record at that citation is *Brown v. Board of Education*. The

names share "board" and "education", so the verdict would be GROUNDED. The receipt would name *Brown*, and a person who read it would see the mismatch, but the verdict would admit.

- *United States v. Smith*, 905 F. Supp. 2d 121. The record is *United States of America v. Iss Marine Services, Inc.* The names share "united" and "states": GROUNDED.

The same holds for any caption whose shared word is a government party (United States, State, People, Commonwealth), a generic term (Estate, Board, County), or, in an airline case, the airline. All six fabricated decisions in *Mata* name an airline. The property this paper's run relies on, that no fabrication reaches GROUNDED, held because none of the five *Mata* citations lands on a real case that shares a party word with its claimed caption. It is a property of those inputs, not of the contract. *Fix:* compare the adverse, non-governmental party; ignore sovereign and generic words; require agreement on court and year where the brief gives them; and hold when the comparison is weak.

7.2 False rejections of real law written differently

The floor's own stated rule is that "unresolved-but-possibly-real" law is UNVERIFIABLE, "never NOT_FOUND" (README.md, lines 25–27). Two paths break it.

- **Names.** A single-reference lookup by name returns NOT_FOUND unless one of the top eight records has exactly the same name after case and punctuation are normalised (floor.py, lines 109–116). Standard citation practice abbreviates case names, and a short form such as *Brown v. Bd. of Educ.* normalises to a string no record carries. By the code, Brown would be reported as fabricated. The committed tests use only full, exact names (tests/test_floor_live.py, lines 60–71).
- **Citations.** Citations are compared after case and spacing are normalised, and nothing else (floor.py, lines 38–39, 106–108 and 154–161). A real citation written with a pin cite ("384 U.S. 436, 444"), with its parenthetical ("384 U.S. 436 (1966)"), or with the reporter closed up ("905 F.Supp.2d 121") matches no record's citation exactly. In a covered reporter the verdict is NOT_FOUND, a false accusation. Our inputs were normalised by hand (§4); a real brief's are not. *Miller* appears in the record as "174 F.3d 366, 371–72".

Fix: parse every citation into volume, reporter and first page with a citation parser such as eyecite [10] before the lookup; hold any input that does not parse, or parses more than one way; and make a miss on a name alone UNVERIFIABLE, never NOT_FOUND.

7.3 The good-law veto is off the tested path

The pair check never consults treatment (floor.py, lines 141–166). The veto is wired only into single-reference resolution (lines 118–121). Its free-tier signal is read from the treated case's own record (lines 72–82). Negative treatment is recorded in *later* opinions, not in the treated case's caption, so this signal will miss nearly all of it. By inspection it can also misfire on a caption that contains a treatment word in another sense; a party named "Vacation …", for instance, would trigger it. Neither the run nor any test produces NOT_GOOD_LAW. We therefore describe the veto as wired but unmeasured, with free coverage close to none. "Partial" would overstate it.

7.4 Retrieval, coverage and silent failures

- **Top results only.** The pair check examines the first five results of its query and the single lookup the first eight (floor.py, lines 96 and 141). The pair check compares the name with the first record that carries the citation (line 162). Where two records share a page, as the two orders at 92 F.3d 1074 do (¶ 29(c)), a correct claim naming the second record would be MISCITED, and a fabrication sharing a word with the first would be admitted.

- **Coverage is assumed.** The covered set rests on the comment that the database covers those reporters "comprehensively" (lines 31–35). We did not measure it. One opinion missing from a covered reporter turns a real citation to it into NOT_FOUND.
- **Wrong but well-formed answers.** A response with an empty result set returned in error, or with a renamed citation field, would be read as "no record" (`courtlistener.py`, line 27) and would produce NOT_FOUND across the covered reporters (§3.3). Canary citations are the guard, and they must be part of every batch.

7.5 Our own descriptions, corrected

The audit also found places where the lane describes itself more generously than its code or output supports. The demonstration script's docstring says it catches "every one" of the *Mata* fabrications (`dogfood_mata.py`, line 5), while its own score line reports three caught and two held. The script's input list is described as the citations fabricated "per the court's sanctions order" (line 24), but it holds five of the six decisions and none of the authorities in ¶ 29. The README reports "6 gates green" (line 22); that figure predates the second increment, and the test file now has 11. The Admissible Motion paper describes its motion floor as sharing "one engine with the bio and legal admission floors" [22]. The legal floor shares the contract, not the engine (§3). The run's numbers in this paper are taken from the run, not from any of these descriptions.

8 The next experiment

The run answers whether the contract behaves on the inputs it was given. The open question is different: on a real legal task, is any learned component worth adding next to the strongest deterministic verifier we can build? We specify the experiment here, before running it, in the style of our Inference Placement and Irrecoverability programmes [24, 25]. There, the strongest deterministic method is built first, the learned lane is handed every advantage, and the bar is fixed before any outcome is seen. Across those programmes the learned lane has so far not earned a seat (in *Irrecoverability*, zero of 160 world-by-dimension cells). We expect the same here and will report it either way.

Task. For each authority in a filing, given as a case name, a citation and the proposition it is cited for, decide *admit*, *reject* or *hold*, with a receipt.

Test set, fixed before the first run. The set is directed at failures: every audit finding of §7 contributes inputs built to trigger it.

(i) All 18 authorities of the *Mata* record (Table 1), labelled from the court's findings. The seven fabricated or not-as-cited citations that were not run come first. The two further decisions the court's order of 11 April required (¶ 13) join only after their status is established from the record.

(ii) Real authorities drawn from the tables of authorities of published opinions, labelled against a source other than CourtListener, so that the floor's database is never its own ground truth. The Caselaw Access Project [4] is one candidate.

(iii) Perturbations for each audit finding: fabricated captions that share a party word with the real case at the citation (§7.1); abbreviated case names, pin cites, parentheticals and closed-up reporters on real citations (§7.2); pages shared by two records (§7.4); Westlaw and neutral citations, both real and fabricated; and pages that fall inside another opinion.

(iv) Canary citations in every batch. A batch is void if any canary fails to ground (§3.3).

Deterministic competitor, built first. The competitor parses citations with eyecite [10] and looks each one up exactly. It then applies four checks: page-range containment, to name the opinion a page falls in, as the court did; a party-aware name comparison; agreement of court and year; and the coverage rule, with abstention on every error. Where a token is available, CourtListener's citation-lookup API [8] runs alongside as a second, independent lookup.

Learned lane, confined. A local model sees the same evidence, including the opinion text for class E. It may move an authority from *admit* to *hold*, for example to flag that the opinion does not appear to support the proposition. It may propose a disambiguation, which the deterministic verifier must then re-check. It may never admit, and it may never move an authority out of *reject*. This is the narrowing-only constraint of our VDSG runtime and Fail-First Models, applied to citations [27, 23].

Gate, fixed now. The loss counts three kinds of error on real authorities: false rejections; holds on authorities that do support their propositions; and admissions of authorities that do not (class E). The confined learner can lower the second by proposing disambiguations that the deterministic verifier confirms, and the third by flagging misuse; its flags can also raise the second. It earns a seat only if it meets all three conditions of the gate we inherit from Inference Placement [24]: it reduces that loss by at least 15%; the paired 90% bootstrap confidence interval of the reduction excludes zero; and admissions of fabricated or miscited citations do not rise. One such admission disqualifies the learned lane.

Reporting. This section fixes the task, the competitor, the confinement and the gate in advance [18]. The test set is to be frozen before the first run. Every arm will be reported, including null and negative results.

9 Related work

Legal hallucination. Dahl et al. give systematic evidence on legal hallucinations. When asked specific, verifiable questions about random federal court cases, the models they tested hallucinated between 58% of the time (ChatGPT 4) and 88% (Llama 2), and the models "cannot always predict, or do not always know" when they are hallucinating [11]. In an appendix they ask models whether given cases exist, using fabricated cases built with plausible party names and the right reporter. Two models tended to assert the existence of "any case—real or fake", and two tended to deny it [11, App. E.1]. Both biases are failure modes a self-check cannot rule out. The floor is aimed at both: it admits nothing a database record does not back, and it rejects nothing whose absence the database cannot vouch for. Magesh et al. ran a pre-registered evaluation of commercial legal research tools built on retrieval-augmented generation. The preprint reports that each hallucinated between 17% and 33% of the time, and it distinguishes correctness from groundedness [16]. Their *misgrounded* category is our class E, the class an existence check admits by design (§6). Where their unit is a research answer, ours is a single citation, and our claim is correspondingly narrower. LegalBench measures legal reasoning across 162 tasks [13]; citation existence is not its subject.

Fabricated references outside law. Walters and Wilder had GPT-3.5 and GPT-4 write 84 short literature reviews, containing 636 references in all. They found 55% of GPT-3.5's references fabricated, against 18% of GPT-4's. Among the real references, 43% and 24% respectively contained substantive citation errors [21], the scholarly analogue of MISCITED. Agrawal et al. show that hallucinated references can often be detected without any external resource, by asking a model consistency questions about the reference it produced [2]. That line of work improves the proposer's self-knowledge. Ours puts an external record between the proposer and the filing. The two are complementary, but only the second produces a receipt.

Citation infrastructure. CourtListener, from the Free Law Project, provides the search API the floor uses [9]. It also provides a citation-lookup API that "can be useful as a guardrail to help prevent hallucinated citations". That API needs a token, is throttled to 60 valid citations per minute, parses citations from free text with eyecite, and reports a valid citation it cannot find as 404 and an ambiguous one as 300, over a database its documentation puts at 18,128,182 citations [8]. That service is the closest prior art for the existence check, and a deployed floor should use it. Its documentation describes matching by volume, reporter and page, and we found no description of a check of the case name against the citation. The name-to-citation comparison (MISCITED), the coverage rule and the admission contract are what the floor adds on top. Eyecite parses legal citations [10] and is the natural fix for the normalisation failures of §7.2. The Caselaw Access Project offers free, public access to "over 6.5 million decisions published by state and federal courts" [4]. It is a second free source against which CourtListener's coverage could be tested. Commercial citators such as Shepard's and KeyCite provide the treatment history the floor lacks (§3.5). We did not evaluate them or any other commercial tool.

Attribution in language generation. Retrieval-augmented generation conditions a generator on retrieved text but leaves the generator the author of the output [15]. Attribution to identified sources, AIS [19], ALCE's citation-quality metrics [12] and FActScore's atomic-fact precision [17] measure whether generated text is supported by what it cites. They evaluate support, which is class E. The floor enforces existence and identity, and a claim that fails either one is never admitted. Kalai and Vempala show that calibrated language models must hallucinate facts that are arbitrary with respect to their training data at a rate tied to how rarely such facts appear [14]. The volume and page of a rarely cited opinion look like such facts. This is consistent with the view that citations should be checked, not recalled.

Abstention and runtime monitors. The three-way verdict is the reject option of classical decision theory [5], applied with an asymmetric cost that the legal setting supplies: calling real law fabricated is itself a harm. Structurally, the floor is a runtime monitor between a proposer and trusted state, as in Simplex architectures and shielding [20, 3]. Here the protected state is a filing, not an actuator.

Our earlier work. The admission pattern comes from our science floor. *The Verification Floor* measured zero false acceptances in 1,737 adversarial cases across a 14× range of model sizes [28]. The legal floor has no adversarial suite of that kind yet, and §7.1 shows why it needs one before any rate is claimed. *Peel* makes a typed store the only author of facts and demotes models to proposers or renderers [26]; here a public case database plays the store. *We Tried to Train It In* reports that training grounding into 3-billion-parameter models failed, while a deterministic resolver outside the model closed the gap [29]. The legal floor makes the same placement decision without first trying to train it in. *Admissible Motion* names the legal floor as one instance of the cross-domain pattern [22]. The method of §8 comes from *Inference Placement* and *The Irrecoverability Boundary* [24, 25], and its confinement of the learner from VDSG and *Fail-First Models* [27, 23].

10 Limitations

Status. Research prototype. Not legal advice and not a citator. A GROUNDED verdict means that a citation exists in one free database and names the claimed case. It does not mean the case is good law, that it supports anything, or that it may be relied on. Every figure is ours, and none has been replicated by a third party.

Sample size. Five fabricated citations and three real ones, in one run. We report no rates. Zero false admissions in five is compatible with a false-admission rate as high as 52% (upper end of the two-sided 95% Clopper–Pearson interval [6]), and zero false rejections in three with a false-rejection rate as high as 71%. Neither bound says anything useful, and the audit exhibits inputs on which the floor would fail.

Easy controls. The three real citations are among the most-cited decisions in United States law, in the reporter the database covers best. They show that the path works. They do not estimate how often real law is falsely rejected. The audit shows two ways that can happen (§7.2).

Hand-normalised inputs. The run's citations were stripped to bare form by hand. On citations as a brief writes them, the floor as built would reject some real law (§7.2).

Partial coverage of the record. Five of the eighteen authorities in the opinion were run. *Shaboon*, the six fabricated or not-as-cited authorities inside the fake *Varghese*, and the six misused real decisions were not. For those, Table 4 gives the contract's prescription, not a result.

Unmeasured assumptions. Rejection is sound only under coverage, retrieval and canonical input (Proposition 1). None of the three is measured here.

Audit findings not executed. The failure paths of §7 are properties of the code, established by reading it. Each could be confirmed with a few inputs, and none has been.

Tests not re-run. The lane's 11 committed tests were not run for this paper. Their last recorded session left no failures, which a network-less, all-skipped session would also do (§4).

Dated evidence. CourtListener changes. A record can be added, corrected, merged or removed, and a verdict from 27 September 2026 may not be reproducible later. Receipts are links, not archived copies, and they are not signed. We archived the run's output and hashes (Appendix B), not the records behind the links.

Scope. United States case law only; the covered set is federal and Supreme Court reporters only. No statutes, regulations, secondary sources or foreign law. No evaluation of any commercial tool.

11 Conclusion

A generated citation looks like a real one, and asking the generator whether it is real does not help: in *Mata*, the generator said it was. What helps is a record. The citation floor turns that observation into a contract. A proposal is admitted only when a public database holds an opinion at the citation under the claimed name. It is rejected when the database shows the citation belongs to another case, or when coverage says nothing is there. It is held whenever the database cannot tell, including whenever the database fails.

On the *Mata* citations it was given, the floor admitted none of five fabrications, rejected three, held two, and grounded three real citations. Its one MISCITED receipt named the same real case the court found. On the full record it reached five of eighteen authorities. It cannot reach six of them by any existence check, because those six are real cases misused. Reading its code, we found inputs on which it would admit a fabrication or reject real law. The first of those failures would break the property the run appears to demonstrate.

The contract is sound in shape and unproven in its details. Its rejections are exactly as good as the coverage assumption behind them; its admissions are only as good as the name comparison. The next

step is to run it on everything the record contains and everything the audit found, beside the strongest deterministic verifier we can build, with the gate for any learned component fixed in advance.

References

[1] *Mata v. Avianca, Inc.*, No. 22-cv-1461 (PKC), Opinion and Order on Sanctions, ECF No. 54 (S.D.N.Y. June 22, 2023), 43 pp. Obtained from the RECAP archive, https://storage.courtlistener.com/recap/gov.uscourts.nysd.575368/gov.uscourts.nysd.575368.54.0.pdf (SHA-256 c6fd7073…d932a7dff; full hash in Appendix B). Paragraph numbers (¶) refer to its findings of fact.

[2] A. Agrawal, M. Suzgun, L. Mackey, A. T. Kalai. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 912–928, 2024. doi:10.18653/v1/2024.findings-eacl.62.

[3] M. Alshiekh, R. Bloem, R. Ehlers, B. Könighofer, S. Niekum, U. Topcu. Safe reinforcement learning via shielding. In *Proc. AAAI Conference on Artificial Intelligence* 32(1), 2018. doi:10.1609/aaai.v32i1.11797.

[4] Harvard Law School Library Innovation Lab. Caselaw Access Project. https://case.law; project page https://lil.law.harvard.edu/projects/caselaw-access-project/, accessed 27 September 2026.

[5] C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* 16(1):41–46, 1970. doi:10.1109/TIT.1970.1054406.

[6] C. J. Clopper, E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26(4):404–413, 1934. doi:10.1093/biomet/26.4.404.

[7] Free Law Project. CourtListener: advanced search and query techniques (the `citation` field; quoted phrase queries). https://wiki.free.law/c/courtlistener/help/search/advanced-search-and-query-techniques, accessed 27 September 2026.

[8] Free Law Project. CourtListener Citation Lookup and Verification API, REST v4. https://wiki.free.law/c/courtlistener/help/api/rest/v4/citation-lookup, accessed 27 September 2026.

[9] Free Law Project. CourtListener Search API, REST v4. https://wiki.free.law/c/courtlistener/help/api/rest/v4/search, accessed 27 September 2026.

[10] J. Cushman, M. Dahl, M. Lissner. eyecite: a tool for parsing legal citations. *Journal of Open Source Software* 6(66):3617, 2021. doi:10.21105/joss.03617.

[11] M. Dahl, V. Magesh, M. Suzgun, D. E. Ho. Large legal fictions: profiling legal hallucinations in large language models. *Journal of Legal Analysis* 16(1):64–93, 2024. doi:10.1093/jla/laae003. arXiv:2401.01301 (appendix cited from v2).

[12] T. Gao, H. Yen, J. Yu, D. Chen. Enabling large language models to generate text with citations. In *Proc. EMNLP 2023*, pp. 6465–6488. doi:10.18653/v1/2023.emnlp-main.398.

[13] N. Guha, J. Nyarko, D. E. Ho, C. Ré, et al. LegalBench: a collaboratively built benchmark for measuring legal reasoning in large language models. arXiv:2308.11462, 2023.

[14] A. T. Kalai, S. S. Vempala. Calibrated language models must hallucinate. In *Proc. 56th ACM Symposium on Theory of Computing (STOC)*, pp. 160–171, 2024. arXiv:2311.14648.

[15] P. Lewis, E. Perez, A. Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2005.11401.

[16] V. Magesh, F. Surani, M. Dahl, M. Suzgun, C. D. Manning, D. E. Ho. Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies* 22(2):216–242, 2025. doi:10.1111/jels.12413. Figures and quotations are from the preprint, arXiv:2405.20362v1.

[17] S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi. FActScore: fine-grained atomic evaluation of factual precision in long form text generation. In *Proc. EMNLP 2023*, pp. 12076–12100. doi:10.18653/v1/2023.emnlp-main.741.

[18] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, D. T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115(11):2600–2606, 2018. doi:10.1073/pnas.1708274114.

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

[20] L. Sha. Using simplicity to control complexity. *IEEE Software* 18(4):20–28, 2001. doi:10.1109/MS.2001.936213.

[21] W. H. Walters, E. I. Wilder. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports* 13:14045, 2023. doi:10.1038/s41598-023-41032-5.

Our earlier papers (Perslis Research, September 2026)

[22] Admissible Motion: runtime safety as an admission-control problem, not a model-capability problem. research.perslis.com/motion.

[23] Fail-First Models: failure becomes structure, structure changes the next attempt. research.perslis.com/fail-first.

[24] Inference Placement: where learned inference earns authority in a provenance-constrained symbolic system. research.perslis.com/inference-placement.

[25] The Irrecoverability Boundary: where learned inference cannot recover what deterministic computation cannot reach. research.perslis.com/irrecoverability.

[26] Peel: structural hallucination prevention for offline AAC through symbolic fact authorship. research.perslis.com/peel.

[27] VDSG: a commanded admission-control runtime for autonomous agents. research.perslis.com/vdsg.

[28] The Verification Floor: scientific integrity that does not degrade with model scale. research.perslis.com/verification-floor.

[29] We Tried to Train It In: negative results on teaching small models to ground facts. research.perslis.com/training-grounding.

A The run, verbatim

Below is the demonstration script's complete output from the single live run. We removed the ANSI colour codes. For typesetting only, we dropped the title's emoji, replaced the hooked arrow before each receipt with "->", and wrapped long lines. The unedited file is in `evidence/`, with its hash in Appendix B. The script's headings ("Fabricated by ChatGPT (should be CAUGHT)") are its own words. Two of its five "should be caught" inputs were held, not caught.

Transcript of the single live run, mata-run-20260927T090825Z.txt

```
THE LEGAL TEST — Mata v. Avianca fabricated citations vs the floor

Fabricated by ChatGPT (should be CAUGHT):
  NOT_FOUND    Varghese v. China Southern Airlines Co., Ltd., 925 F.3d 1339
    no case carries this mainstream-reporter citation — fabricated, REJECTED
  MISCITED     Petersen v. Iran Air, 905 F. Supp. 2d 121
    citation is real but names "United States of America v. Iss Marine Services, Inc.", NOT
        "Petersen v. Iran Air" — REJECTED
    -> United States of America v. Iss Marine Services, Inc. — 905 F. Supp. 2d 121, 84 Fed.
        R. Serv. 3d 384, 2012 WL 5873682, 2012 U.S. Dist. LEXIS 166088 — https://www.courtli
        stener.com/opinion/2661490/united-states-of-america-v-iss-marine-services-inc/
  NOT_FOUND    Miller v. United Airlines, Inc., 174 F.3d 366
    no case carries this mainstream-reporter citation — fabricated, REJECTED
  UNVERIFIABLE Martinez v. Delta Airlines, Inc., 2019 WL 4639462
    citation not in the free corpus (e.g. Westlaw/unpublished) — cannot verify
  UNVERIFIABLE Estate of Durden v. KLM Royal Dutch Airlines, 2017 WL 2418825
    citation not in the free corpus (e.g. Westlaw/unpublished) — cannot verify

Real law (should be GROUNDED):
  GROUNDED     Miranda v. Arizona, 384 U.S. 436
    -> Miranda v. Arizona — 16 L. Ed. 2d 694, 86 S. Ct. 1602, 384 U.S. 436, 1966 U.S. LEXIS
        2817, 10 Ohio Misc. 9, 36 Ohio Op. 2d 237, 10 A.L.R. 3d 974 —
        https://www.courtlistener.com/opinion/107252/miranda-v-arizona/
  GROUNDED     Brown v. Board of Education, 347 U.S. 483
    -> Brown v. Board of Education — 347 U.S. 483, 74 S. Ct. 686, 1954 U.S. LEXIS 2094, 38
        A.L.R. 2d 1180, 53 Ohio Op. 326, 98 L. Ed. 873 —
        https://www.courtlistener.com/opinion/105221/brown-v-board-of-education/
  GROUNDED     Gideon v. Wainwright, 372 U.S. 335
    -> Gideon v. Wainwright — 372 U.S. 335 —
        https://www.courtlistener.com/opinion/8954562/gideon-v-wainwright/

SCORE
  fabrications caught (NOT_FOUND/MISCITED): 3/5   (+2 UNVERIFIABLE = free-data gap, honestly
      abstained)
  fabrications that SLIPPED THROUGH:        0
  real law grounded with receipts:         3/3
  real law FALSELY rejected:                0

  FLOOR HELD — no fabrication reached a verified state, no real law falsely rejected.
```

B Reproduction, evidence and code map

Evidence files (all under `evidence/` in this paper's folder).

- The run's raw output, `mata-run-20260927T090825Z.txt`, with SHA-256 `8be56a5f725ad28a66791a6c91af1f05be8978db4d518c11c2d02c30eb1f86d0`.
  Its `.meta.txt` records the command, start and end times (09:08:25Z and 09:08:33Z), exit code 0, Python 3.9.6, the lane's repository head, the last commit to touch the lane (`533f783`), an empty diff against that commit, and that no API token was set.
- `sha256-lane-files.txt`: hashes of the eight lane files read for this paper (README, demonstration script, the five package files, the test file).
- The court's opinion as text, `mata-opinion-ecf54.txt`, extracted with `pdftotext -layout` from the source PDF (43 pages, 2,341,033 bytes) with SHA-256 `c6fd707305765357003654c5ed87426e3605e1ee6bdd3d7cb9ff1a7d932a7dff`.
  The `.meta.txt` gives its URL.
- `pytest-cache-snapshot/`: the lane's test cache as of 23 September (11 test identifiers; an empty failure list), copied with hashes, because the next test session would overwrite it.

Rebuilding every number, table and figure (read-only; no network).

```
python3 -B analysis/parse_run.py       # recount the run
python3 -B analysis/record_table.py    # pin the record to opinion lines
python3 -B analysis/make_tables.py     # Tables 1 and 3, TeX and HTML
python3 -B analysis/make_transcript.py # Appendix A
python3 -B analysis/bounds.py          # Section 10 bounds
python3 -B web-build/export_figs.py    # Figures 1 and 2
```

Each script aborts if its input disagrees with what it expects. The live run itself was made once, with the command recorded in the `.meta.txt`. Repeating it would query CourtListener again, and its answers may have changed.

Code map. Lines refer to commit `533f783` of the lane.

- *Exact citation query*: `legal_floor/connectors/courtlistener.py`, lines 51–58.
- *Throttling and retries*: `legal_floor/http.py`, lines 16–18 (retryable codes; four attempts), 28–35 (one-second spacing per host), 38–56 (backoff, `Retry-After`, 30-second cap, fail-fast), 59–70 (non-JSON becomes a source error).
- *Source error becomes* `UNVERIFIABLE` (Proposition 2): `legal_floor/floor.py`, lines 100–104 and 149–152.
- *Coverage rule*: `floor.py`, lines 31–35 and 156–161.
- *Name comparison*: `floor.py`, lines 29–30 and 129–138; *first record only*: line 162.
- *Treatment veto*: `floor.py`, lines 28, 72–82 and 118–121; absent from `verify_citation`, lines 141–166.
- *Unused token path*: `courtlistener.py`, lines 61–86 (defined, never called by `floor.py`).