CORTEGA LABS

Verification Controls for Legal AI

Research report · October 2026

This edition presents selected standalone-model results from the October study. These model measurements are separate from the Cortega Legal system result of 91.4% of criteria met on 1,743 Harvey LAB tasks.

Citation benchmark results and implementation requirements for legal research products

CORTEGA LABS | Research report | October 2026

Executive summary

Legal research products need to establish whether cited authorities exist, whether references identify the intended cases, and whether the opinions support the claims made in an answer. These are different checks. A response can cite a real case while quoting it incorrectly or attributing a proposition the opinion does not support.

Cortega's October 2026 evaluation compared original model answers with answers processed through its AI Verifier. Eight models received the same 100 U.S. legal research questions. Citations were checked against a local mirror of CourtListener records, and models could revise an answer up to twice. Across all models, citations classified as wrong fell from 1,045 to 163. The share of wrong citations declined for every model. [1, 2]

The reduction needs to be assessed alongside research coverage. Revised answers contained more citations rated relevant and supported for four models, fewer for three, and the same number for one. Total citation counts changed as answers were rewritten. Some models returned empty answers after reaching their token limits; citation-level results exclude those answers.

The evaluation used automated checks and AI judges. No lawyer reviewed the results. Some references could not be checked against the available records and were excluded from the wrong-citation count. The test covers one run of a U.S. question set. It does not establish that a revised answer is complete, legally sound or suitable for a specific matter.

Implications for product teams

A verification feature should distinguish confirmed errors, weak support and unresolved references. Removing an incorrect citation can improve accuracy while reducing useful research. Product evaluation should therefore measure answer completion, retained supported authorities, reviewer acceptance and the cost of resolving flagged claims.

This paper documents the test method, results and limitations, then describes implementation requirements for citation verification in a legal research application. The recommended workflow includes traceable source references, bounded correction and explicit review of unresolved results. Requirements for checking subsequent case treatment, jurisdictional applicability and matter-specific facts are discussed as additional work; the benchmark did not establish those capabilities.

Visual guide: verification and evidence

The chart reports aggregate results under the recorded study conditions. See the methodology and limitations in this paper before applying the findings to a product decision.

Public benchmark overview

1. Verification scope and error types

Citation verification combines record matching with assessment of the relationship between an answer and an opinion. Record matching can identify whether a case exists at a supplied reference. Assessing the proposition attributed to that case requires locating and interpreting the relevant text. An application should report these outcomes separately so a reviewer can understand what passed and what remains unresolved.

Check	Error assessed in this evaluation
Case existence	No case in the available court records has the supplied reference.
Case identity	The case name does not match the case at that reference.
Citation reference	The supplied first page is not the first page of the case.
Pinpoint reference	The page cited for a proposition is outside the case’s pages.
Quotation	The quoted words are absent from the opinion.
Opinion support	The answer misrepresents what the opinion says.

Table 1. Error categories used in the October citation benchmark. A citation can contain more than one error but is counted once in the wrong-citation total. [1, 3]

Relevance and support

The test also assessed whether a cited case was relevant to the question and whether the opinion supported its use in the answer. The published data distinguish citations rated good, weak, inaccurate, hallucinated and uncheckable. In the aggregate tables in this paper, “wrong” means inaccurate plus hallucinated. Weak and uncheckable citations are not included in that count.

Additional requirements

An existing case and a supported quotation do not establish that the case remains authoritative, governs the jurisdiction, or applies to the matter’s facts. Those questions need additional sources and review. Similarly, checking the citations that appear in an answer does not establish that the answer includes every material authority. A complete research product needs to define how it evaluates omitted authorities and limits its claims when it lacks the required source coverage.

2. Benchmark methodology

Each model received 100 U.S. legal research questions: 30 landmark-case questions, 35 legal-doctrine questions, 20 drafting tasks and 15 questions concerning less common cases. All eight models used the same question set. The dataset was designed to include both common authorities and less familiar questions where fabricated references may occur. [1]

Initial answer and reference checks

The original answer was checked against a local mirror of CourtListener court records. Checks assessed nonexistent cases, mismatched names, incorrect first pages, pinpoint references outside the opinion, absent quotations and misrepresentation of opinion content. AI judges assessed relevance to the question and support from the opinion. This combines record checks with judgment-based assessments; the results should not be described as entirely deterministic.

Revision process

Each answer then went through Cortega's AI Verifier. A model could rewrite its answer up to twice, and the rewritten answer was checked again. This run also asked models to improve research coverage using a target set for each question. The comparison therefore measures a verification-and-revision process, including the coverage instruction, rather than a classifier applied to unchanged text.

Counting rules

The wrong-citation count combines inaccurate and hallucinated citations. A citation is counted once as wrong even if multiple error categories apply. The percentage denominator is the total number of citations in that condition. Rewriting can remove, replace or add citations, so the baseline and revised percentages use different denominators.

The full original study contained 757 nonempty answers across eight models; this edition's CSV contains the selected models shown below. Each row records citation totals and category counts before and after revision. The remaining 43 of the 800 assigned model-question tasks produced empty answers, as reported in the source analysis. They do not contribute citations to the error-rate calculation. [1, 2]

Review and operational measurements

3. Citation results

Model	Answered / 100	Baseline wrong	Revised wrong
Claude Opus 5.5	100	187 (11.8%)	23 (1.7%)
GPT-6 Sol	100	39 (7.6%)	2 (0.4%)
Gemini 3.8 Flash	100	77 (15.6%)	8 (2.1%)
DeepSeek V4.1 Flash	94	94 (7.6%)	16 (1.8%)
Kimi K3	97	160 (13.5%)	51 (5.0%)
MiniMax M3	100	183 (22.8%)	20 (3.4%)

Table 2. Wrong-citation counts and share of citations, before and after verification and revision. "Answered" reports nonempty answers in the source dataset. Each model received 100 questions. Model names are the labels recorded in the benchmark. [1, 2]

Across the full original study, wrong citations declined by 882, from 1,045 to 163, an 84.4% reduction in the observed count. This is not an 84.4-percentage-point improvement and is not a comparison against a fixed number of citations. The revised answers contain a different total set of references.

Examples from the results

MiniMax M3's wrong-citation count fell from 183 to 20. Its hallucinated-category count, which includes invented or mismatched citations in this evaluation, fell from 66 to 1. GPT-6 Sol's wrong-citation count fell from 39 to 2. These results show improvement on the assigned question set; they do not establish error rates for a different workload.

Model comparison

4. Research coverage and completion

Model	All citations before / after	Supported citations before / after	Empty answers / 100
Claude Opus 5.5	1588 / 1344	900 / 1048	0
GPT-6 Sol	512 / 496	386 / 432	0
Gemini 3.8 Flash	495 / 374	292 / 292	0
DeepSeek V4.1 Flash	1239 / 877	743 / 679	6
Kimi K3	1181 / 1013	647 / 745	3
MiniMax M3	804 / 588	348 / 399	0

Table 3. Citation volume, citations rated relevant and supported, and empty answers. “Supported” reproduces the dataset’s good category; these ratings were automated. Counts measure citation occurrences, not necessarily distinct authorities. [1, 2]

Interpreting answers with no detected error

GPT-6 Sol produced 98 answers with no detected wrong citation after verification, compared with 82 before. This measure permits weak and uncheckable citations. It does not mean 98 answers were complete or legally correct. An application should avoid labelling such answers “verified” without describing the checks covered by that status. [1]

Handling incomplete results

An empty answer, an uncheckable source and a confirmed citation error require different responses. An empty answer needs a task-completion fallback. An uncheckable source needs another source or review. A confirmed error needs correction or removal. Combining these outcomes in one success score can reward a system for returning less research rather than delivering a usable answer.

For product reporting, retain all assigned questions in completion measures, including questions with no output. Report citation-level accuracy separately. This allows a reviewer to assess whether an apparent accuracy improvement came with more unanswered work or less supported material.

5. Evaluation limits

Question-set and source coverage

The evaluation covers U.S. legal research and a single run of 100 questions per model. It does not establish performance for other jurisdictions, practice areas or customer-specific research tasks. The CourtListener mirror defines the records available to the checks. A failed lookup must be distinguished from a confirmed nonexistent case when the collection may be incomplete.

Some references could not be checked and were excluded from wrong-citation counts. Those references must remain visible in product reporting. Otherwise an answer with unresolved citations could appear equivalent to one

whose sources were found and assessed.

Automated assessment

No lawyer reviewed the results. Automated assessment can misjudge whether an opinion supports a proposition, whether a quotation matches its intended passage, or whether an authority is relevant to a question. The involvement of tested models in assessment adds a potential source of bias. A follow-up evaluation should use practitioner review of a sample spanning errors, clean results and unresolved references.

Changes introduced by revision

The procedure allowed up to two rewrites and also requested improved coverage. Baseline and revised answers therefore differ in wording, citation count and research content. The observed improvement cannot be attributed solely to identifying nonexistent cases. Product tests should preserve the original and each revision so reviewers can determine what changed and whether the change addressed the stated issue.

Operational evidence not collected

The benchmark did not measure verifier cost or added latency. It also did not measure lawyer acceptance, time spent reviewing flags, or the cost of correcting unresolved answers. Those measurements are necessary to decide whether verification improves a production workflow. The security benchmark's latency and cost figures should not be reused as estimates for legal verification.

Conclusions supported by this test

The verification-and-revision procedure reduced detected citation errors across all eight models on this question set. Coverage outcomes were mixed, and some assigned questions remained unanswered. The evidence supports further workflow evaluation of this procedure; it does not establish complete research, current authority or matter-specific legal correctness.

6. Product implementation requirements

Cortega's legal-verifier product materials describe checking citations and opinion support and returning diagnostics for correction. The requirements below explain how a legal research product can expose those checks to its users. They are implementation recommendations, not additional capabilities established by the benchmark. [4]

Source and claim traceability

Associate each material proposition with its citation and the opinion passage used to assess it. Preserve the original answer, resolved case identity, source location, retrieval time and verification outcome. A reviewer should be able to inspect the passage without reconstructing the model prompt. Store identifiers and versions that allow the checked evidence to be distinguished from a later source update.

Result categories

Return specific outcomes such as missing record, identity mismatch, unsupported quotation, disputed proposition and unresolved reference. Keep relevance and support separate from record existence. Use "no detected error in the checks performed" where appropriate instead of a broad assertion that an answer is legally verified.

Correction and review

Bound correction attempts by a configured retry count, time limit and cost limit. Recheck the full revised answer because a rewrite may introduce new references or alter another claim. If a material issue remains unresolved, preserve it for review rather than removing the diagnostic while delivering the answer. Show the difference between the original and revised text when a reviewer needs to approve the change.

Coverage and authority

Evaluate whether revision preserves the relevant authorities and the response needed for the question. If a product claims to check subsequent treatment or jurisdictional applicability, connect those checks to appropriate records and test them separately. Neither source existence nor opinion support supplies that evidence on its own.

Access and delivery

When matter documents are involved, enforce the user's access to the matter and apply data restrictions to model providers, verification services and logs. Define which outputs require practitioner approval before delivery to a client or incorporation into an authoritative document. These workflow controls were not measured by the citation benchmark.

Design principle: layers of legal evidence

A citation that resolves to a real case can still identify the wrong authority, quote the wrong passage or support a different proposition. A supported proposition can still be unsuitable for the jurisdiction or legal issue. Verification therefore needs explicit result categories and a record of the evidence supporting each check.

Cortega's verification approach distinguishes record lookup from contextual support checks. That separation helps users understand what was established and what still needs review. The benchmark evaluates a defined subset of these requirements; current treatment and practitioner judgment require additional evidence.

Legal verification approach

7. Deployment evaluation and references

Select one research workflow and assemble representative questions with practitioner-defined expected authorities and material propositions. Include less common cases, ambiguous citations, quotations, unresolved sources and questions for which the correct response is that the available evidence is insufficient. Keep a held-out set and repeat runs to assess variation.

Measure	Evidence to collect
Citation errors	Confirmed record, quotation and proposition errors, with reviewer adjudication.
Unresolved references	Sources that cannot be checked and the reason each check is incomplete.
Research coverage	Required authorities retained or omitted after revision, by question.
Task completion	Usable answers across all assigned questions, including empty output.
Review effort	Acceptance, edits and time spent resolving flags.

Measure	Evidence to collect
Performance and cost	Added latency, retries and total cost per accepted research task.
Traceability	Original and revised answers, source passages and check versions.

Set release criteria for the selected workflow. Review both missed errors and false alarms; a checker that repeatedly flags valid research adds work. Re-evaluate when the model, source collection, verification rules or correction prompt changes. A product decision should consider the accepted research result and reviewer effort alongside citation accuracy.

References

[1] Cortega. Legal citation errors across eight AI models. October 2026. Method and analysis supplied in the benchmark post folder. Public results: https://cortega.ai/cortega_benchmarks.html#case-citations

[2] Cortega. Per-question citation totals and outcomes before and after verification. https://cortega.ai/cortega_benchmarks.html

[3] Cortega. Original citation-level results and error labels. https://cortega.ai/cortega_benchmarks.html

[4] Cortega. AI Verifier for legal research. Product description. https://cortega.ai/cortega_ai_verifier_legal.html