CORTEGA LABS

Benchmarking AI Models for Legal Applications

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.

Contract analysis, citation accuracy and research coverage

CORTEGA LABS | Research report | October 2026

Executive summary

Cortega's legal benchmark program evaluates two tasks that legal-tech products ask models to perform: extracting provisions from contracts and producing research answers supported by case citations. The tasks have different evidence requirements. Contract findings can be compared with labelled document provisions. Research answers require checking the cited records and whether the opinions support the propositions attributed to them.

We built this program to make model comparison useful to product teams. A legal product needs to know which answers are wrong, whether the model completes the task, how much relevant material it retains, and what changes when verification is applied. A single legal-reasoning score does not provide those distinctions.

The October contract study used 100 contracts from CUAD, an expert-annotated public dataset. Each model received the full contract and the same request for its provisions. A whole contract was counted as correct only when all requested findings matched the labels. Two models each received 74 correct-contract results, while their recorded model cost and response time differed substantially. [1–3]

The citation study assigned the same 100 U.S. research questions to eight models. Answers were assessed before and after verification and revision. Wrong-citation counts fell from 1,045 to 163. Citation shares improved for every model, but supported-citation volume increased for four, decreased for three and stayed unchanged for one. Some assigned questions produced empty answers. [4, 5]

The two studies serve one program. Contract results support document-task selection. Citation results support research-model selection and evaluation of verification. Results should remain separate because an extraction score does not establish citation accuracy, and a citation check does not establish complete contract understanding.

Participation and evidence

We invite model partners to contribute endpoints and legal product teams to contribute tasks and review criteria. Partners gain a common comparison and task-level diagnostics. Users gain a way to examine what a score measures and determine whether it is relevant to their application.

Cortega develops the benchmarks and the evaluated verification product. These are vendor-run evaluations, not independent certification. Basic tests are available in the product. Enhanced benchmarks require a Cortega-assisted engagement on an authorized model or system; test materials, CSV files and raw outputs are not shared. The following sections explain the approach, evidence, limitations and participation.

Visual guide: the legal evaluation

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. Why Cortega built a legal benchmark program

Legal applications require answers tied to evidence. A contract assistant must identify what the supplied agreement says, including missing provisions and the party holding a right. A research assistant must supply authorities that exist and support its answer. A fluent response can fail either task while appearing useful.

We needed an evaluation that preserved those distinctions and let us compare original output with verified output. The approach combines public labelled documents, a task-specific research question set, record checks, automated assessment and per-task reporting. It also retains failures to answer and unresolved references so an improvement in one measure is not presented as improvement in every measure.

Relationship to existing research

LegalBench evaluates a broad range of legal reasoning tasks and includes multiple task structures and domains. It provides valuable domain evidence; it should not be described as a collection of trivial multiple-choice questions. Cortega's program focuses on the particular outputs and comparisons needed for its contract and citation workflows. [6]

CUAD supplies expert annotations for contract analysis. We use a subset of that existing resource rather than claiming to have created independent contract ground truth. CourtListener-derived records support citation lookup. These resources provide different kinds of evidence and should be credited separately. [2, 7]

Relevance to legal products

As a model is incorporated into a document or research workflow, the product must decide how to handle omissions, unsupported claims, unresolved citations and incomplete output. Verification can improve the references in an answer while reducing its useful research content. The decision to add it therefore requires both accuracy and completion measures.

Model hosting, token limits and review procedures also affect the result. A model that performs well on short examples may not complete a full-document task within the supplied limits. The benchmark records these constraints where measured and helps identify what a customer-specific evaluation still needs to test.

What we do not combine

There is no single blended "legal quality" score in this paper. Contract correctness, citation error share, supported research, completion, latency and cost answer different questions. Combining them without an explicit product requirement would make the result harder to interpret and easier to oversell.

2. Contract evaluation: task and scoring

The contract study uses 100 CUAD contracts. CUAD is a public collection of commercial contracts annotated under the supervision of experienced lawyers. Its labels provide a reference for particular provisions; they are not a statement that every possible issue in a contract has been analysed. [2]

Input and task

Each model receives the full contract and a list of requested provisions. It is instructed to answer from the supplied agreement, quote operative wording where possible and state when a requested provision is absent. The runbook records no tools or examples, temperature zero and a 16,000-token response cap. All compared models receive the same task format. [1]

The requested provisions include governing law, liability limitation, insurance, term, termination rights, warranty period, intellectual-property ownership, change of control and restrictive covenants. The findings are assessed against the labelled reference for that document.

Whole-contract correctness

A contract is counted as correct only when every requested finding matches the reference. This is deliberately stricter than averaging correctness across individual provisions. A response with several right findings and one material error is not counted as a fully correct contract.

The metric should not be described as complete legal review. It establishes correctness on the requested, labelled findings. Other risks, drafting quality and matter-specific suitability are outside the task. Results that cannot be scored are recorded separately rather than treated as correct.

Checking the scoring process

The contract assessor was tested against a labelled reference set of 201 fields. It agreed on 99.5%, passed one of 120 wrong findings and rejected none of 81 right findings. An alternative assessor produced different results. This is supporting evidence about the scoring setup, not proof that assessment is infallible on the 100-contract run. [1]

For model partners, the important comparison is within the stated task and scorer configuration. Practitioner review remains useful for contested findings, difficult provision types and applications where a particular error has a higher consequence than others. Public-dataset familiarity may also affect model performance; this study does not rule out training-data overlap.

3. Contract results and model selection

Model	Correct / 100	Unscored	Cost / 100	Seconds
Claude Opus 5.5	74	1	$8.739	9.9
Kimi K3	73	0	$3.905	10.3
MiniMax M3	70	0	$0.375	11.5
GPT-6 Sol	70	0	$2.277	3.2
DeepSeek V4.1 Flash	66	0	$0.457	12.7
Gemini 3.8 Flash	65	0	$1.233	9.6

Table 1. Whole-contract correctness, recorded model token cost per 100 contracts, and average response time. Unscored results remain separate. Model labels and operating measurements reproduce the October dataset. [1, 3]

Reading the comparison

A product team may use the table to shortlist models for a provision-extraction workflow. It should then examine the documents and provision types that failed, and test the candidate on its own contracts. If a model misses termination rights but handles governing-law findings well, that distribution may matter more than its position in the

aggregate table.

Cortega retains document-level outcomes and operating measurements internally. The aggregate table reports whole-contract correctness, model cost and response time. Partners can discuss their model's summary findings with Cortega; document-level outputs and CSV files are not shared. These measurements do not include customer review time or complete service costs. [3]

Cost and deployment

The token-cost comparison is task-specific. It excludes integration, infrastructure, human review and any additional checks required by a product. Hosting can change both cost and response behavior. A private endpoint or local deployment should be evaluated as its own serving configuration.

Select the required provisions, completion limits and review rules before choosing a cheaper model. Test held-out customer contracts and retain the common benchmark as a stable reference.

4. Citation evaluation: task and evidence

The citation study contains 100 questions: 30 landmark-case questions, 35 legal-doctrine questions, 20 drafting tasks and 15 questions about less common cases. Each of eight models receives the same set. The initial run asks models to answer without a search tool, so it measures that configuration rather than retrieval-augmented legal research. [4]

Reference preparation

Cortega constructed the question set and reference cases. Reference identities and relevance were machine-checked before the comparison. Some reference volumes were absent from the available record collection and were flagged. The questions and targets were not independently reviewed by lawyers for this run. The provenance should remain visible alongside the resulting scores.

What is checked

Evidence dimension	Question assessed
Record identity	Does the reference identify a case in the available records, with the matching name and first page?
Pinpoint and quotation	Is the referenced page within the case, and are the quoted words present?
Opinion support	Does the opinion support the proposition attributed to it?
Relevance	Is the authority relevant to the research question?
Availability	Could the source and the required assessment be obtained?

Record matching and opinion assessment are different operations. Finding a case does not establish that it supports the answer. An unresolved lookup also does not prove that a case is fabricated. The dataset distinguishes hallucinated, inaccurate, weak, supported and uncheckable citations. [4, 5, 7]

Original and revised answers

Answers are assessed before and after Cortega's verification-and-revision procedure. Models may revise an answer up to twice, and revisions are reassessed. This run also requests improved research coverage using a target set. The comparison therefore concerns a changed answer, not a classifier applied to unchanged text.

A citation counts once as wrong if it has a hallucinated or inaccurate result, even if several errors are present. Weak and uncheckable references are excluded from the wrong-citation count. Citation totals change when answers are revised, so both the numerator and denominator must be reported.

5. Citation results, coverage and completion

Model	Answers / 100	Original 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 shares, before and after verification and revision. Answer counts report nonempty original answers out of 100 assigned questions per model. [4, 5]

Wrong citations fell from 1,045 to 163 across the eight models. The reduction in observed counts is 84.4%; percentages in the table use each condition's own total citations. The results do not establish that 84.4% of legal errors in a customer workflow would be removed.

Accuracy and coverage together

GPT-6 Sol's wrong citations fell from 39 to 2, while supported citations increased from 386 to 432. It produced 98 revised answers with no detected wrong citation. That measure permits weak and uncheckable references and should not be renamed "98 legally correct answers." [4, 5]

Incomplete answers

Citations are occurrences in generated answers, not necessarily distinct cases. Product selection should consider per-question authorities retained and claims supported rather than interpreting a larger total citation count as inherently better research.

6. Evaluating the evaluator

Users should be able to ask how the scoring process itself was tested. Contract labels, citation record checks and automated opinion assessments provide different evidence. They should not be described as one deterministic legal ground truth.

Supporting validation

The citation checker was separately evaluated on 390 labelled excerpts from LePhantomCite, a public citation-error dataset. The source records report that a tested checker configuration caught 30 of 31 fabricated cases and 11 of 44 misquotes, with false alarms on about 5.2% of correct citations. Detection varied materially by error type. This is supporting validation of the checker, not a score for the eight target models or evidence that every opinion assessment is correct. [4, 8]

The source records also contain probes of opinion-support assessment. Those results do not replace practitioner review of the current question set and generated answers. Automated assessments are not perfectly repeatable, and some assessment models were also tested models.

Limits of the evidence

The citation study had no lawyer review. Some references could not be assessed because of record coverage, including gaps in U.S. Reports volumes. The study covers one U.S. question set and one run per model. It does not establish performance in other jurisdictions, current treatment of authorities, or relevance to a particular client matter.

Revised answers may omit useful material. Verification cost and added latency were not measured per model. Evaluation of live product delivery also remains separate from the saved-answer study. Product teams should not apply the security benchmark's operating measurements to legal verification.

Published evidence and restricted materials

This paper describes the task scope, assessment categories, aggregate results and limitations. Cortega retains document-level results, generated answers and evaluation traces internally. Partners and users are not given access to benchmark CSV files, enhanced test materials or raw outputs.

Enhanced suites can reveal model weaknesses and can be repurposed to develop adversarial tests or attacks against models and enterprise systems. They are restricted to Cortega-assisted evaluations and are not distributed through AI Bench/Red Team. Basic tests remain available in the product.

Trust in the reported results depends on accurate scope, consistent scoring, explicit denominators and disclosure of limitations. It does not imply access to restricted artifacts. Cortega supplies both the evaluation and verification product; external task contributions and practitioner input can strengthen the program, but are not claimed as completed independent validation.

7. Use within legal products

The program supplies task-specific evidence for model selection and verification design. A product should carry the scope of that evidence into its own decisions. A good contract result does not establish case-law research quality, and a reduced citation error share does not establish a complete brief.

Contract workflows

Use provision-level and document-level failures to select candidates for the provisions your application actually needs. Add representative customer contracts, including long agreements, cross-references, exceptions and party-specific rights. Measure review effort and accepted output in addition to correctness. Preserve unanswered and unscored tasks.

Research workflows

Compare original and revised answers on wrong references, unresolved sources, required authorities retained and task completion. Define which claims need practitioner review. A product status should describe the checks performed rather than asserting that an entire answer is legally verified.

For a research application that uses retrieval, tools or a different source collection, benchmark that configuration separately. The October memory-only question run cannot predict how added retrieval changes citation selection or completeness.

Evaluation and operation

Basic benchmark tests are available in the product. The enhanced contract and citation studies described here are not available for self-service through AI Bench/Red Team or as distributed test tools. Contact Cortega to help run the benchmarks on your model or system, with an agreed task scope and authorized access. [1, 4]

Use a benchmark result to guide a controlled application test, then validate the actual reviewer experience, data access, failure handling and response time. A bounded correction process should retain unresolved issues and the original answer, and an empty result needs a completion fallback.

Versioned decisions

Keep the model version, endpoint, prompt, scoring rules and source snapshot with the acceptance decision. Re-evaluate when these change. Public datasets may overlap training data, so customer-specific held-out material provides additional evidence. For confidential matter data, agree permitted storage and review before contributing it to an evaluation.

8. Model-partner participation

We invite model providers and legal product teams to arrange a Cortega-assisted evaluation of contracts, citation research or both. Summary findings identify extraction weaknesses, completion and cost differences, citation-error patterns and the effect of verification. Cortega runs the enhanced tests within the agreed scope; partners do not receive the suites or raw outputs.

The benefit is a common evidence set that helps a model partner discuss a legal use case with product teams. A result can identify a suitable application, a weakness requiring further development or a configuration that needs different limits. Participation should not be presented as a promise of a favorable ranking.

Information for an initial submission

Item	Information to provide
Task scope	Contract extraction, citation research, or both; original-output or verification comparison.
Serving configuration	Model identifier and version, endpoint or deployment arrangement, supported request format.
Limits and access	Secure access, context and output limits, request capacity and pricing assumptions.
Product context	Intended document types, jurisdictions, use of retrieval, and the output users need.
Findings and disclosure	Summary assessment, permitted publication, confidentiality and task contributions. Raw outputs and restricted test materials are not shared.

Agree technical feasibility, operating costs, report scope and publication conditions before running the model. Do not submit API credentials through a public message. A hosted endpoint's result and a local deployment's result are separate observations.

Contributions beyond a model

Legal product teams and practitioners can contribute questions, expected authorities, difficult clauses and review judgments. Such cases should be labelled by provenance and separated from the common comparison set when used for tuning. New jurisdictions or task formats require their own reference coverage and scoring validation.

For an assisted study, agree the summary assessment and how findings will be discussed. Cortega can help investigate an issue or scope a repeat run. CSV files, test prompts, generated answers, source-level evaluation outputs and traces remain internal and are not provided to partners.

Use the Cortega contact page to identify the model and task you want evaluated. Include whether the model is hosted or private and whether the output is intended for contract analysis, research or both. [9]

Design principle: quality has several dimensions

Error reduction is useful when the revised answer still supplies the material needed to complete the task. Removing authorities can reduce the error count while weakening research coverage. A model can also return fewer answers and appear accurate if empty output is excluded without being disclosed.

Cortega's legal program reports contract correctness separately from citation errors, supported material and completion. This gives product and model teams a basis for choosing the relevant tradeoff. Professional review effort remains a separate measurement: neither a low error count nor a low token cost proves that a workflow creates value.

Aggregate legal results

References and evaluation record

[1] Cortega, October contract benchmark runbook and configuration. Legal contract results.

[2] The Atticus Project, CUAD dataset. Expert-annotated source contracts and provisions.

[3] Cortega, October 2026 contract evaluation records. Retained internally; aggregate correctness, cost and timing are reported in this paper.

[4] Cortega, October citation research runbook and saved-answer evaluation. Citation results and limits.

[5] Cortega, October 2026 original and revised citation evaluation records. Retained internally; aggregate results are reported in this paper.

[6] LegalBench, a collaborative benchmark for legal reasoning. Context for complementary evaluation; no endorsement of Cortega is implied.

[7] Free Law Project, citation lookup and verification API documentation. Source-lookup context; Free Law Project did not evaluate Cortega's results.

[8] LePhantomCite dataset. Checker validation used the locally retained 390-excerpt evaluation split. Validation counts are from Cortega's October source records.

[9] Discuss a legal model submission with Cortega.