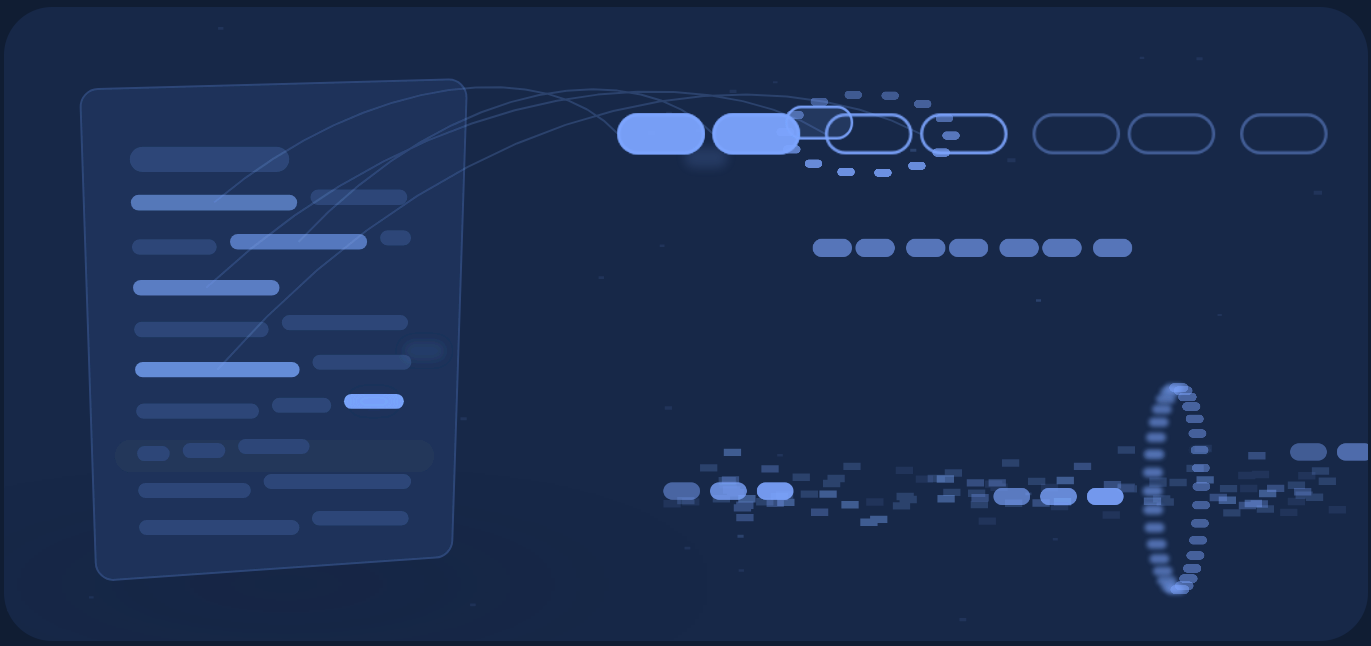
Rillor

RESEARCH REPORT

What benchmarks show about LLM contract review accuracy

Eight benchmarks from 2021 to 2026 test LLM contract review, and scores run high on yes-or-no clause questions and lower on clause extraction and final review.



7 November 2026 Document and regulation intelligence rillor.com/insights/llm-contract-review-accuracy-benchmarks

Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

- Contract review accuracy for language models tracks the task a benchmark sets. On contracts from one public dataset, three commercial models averaged 88.1% to 92.9% balanced accuracy on yes-or-no questions about single clauses in 2023 [3]. The best of 19 model configurations reached 0.644 F1 returning clause text from whole contracts in 2025 [13]. The best model in the dataset's 2021 paper reached 44.0% precision at 80% recall [1], and the best model in a 2026 final-review test found errors at 0.750 macro recall [15].
- The same weak points recur: rare clause categories that score near zero [13], exceptions stated far from their rule [4], drafting errors that need context [15] and long merger agreements [5, 18].
- Absent clauses dominate CUAD's test set, which later benchmarks reuse. About 70% of the questions in CUAD's published test file have no annotated answer, and one public model returned text for 45.4% of them [17].
- Comparisons with lawyers are few and small. In a February 2025 industry study, the top tool scored 75.1 on data extraction against 71.1 for a lawyer control group, and lawyers led on redlining [21]. Two other studies were run by vendors, on ten contracts and on five NDAs [20, 25].
- Several sources contain internal inconsistencies, reported here as found [13, 15, 21]. The findings point to a buyer's test built from the buyer's own contracts that counts misses and absent clauses and reports results by clause type, in line with NIST guidance on realistic test sets [11, 17].

Background

What contract review asks of a model

Contract review covers several jobs. A reviewer locates the clauses that matter, decides what kind each one is, interprets what it requires and checks the document for internal errors. The authors of CUAD, a public contract dataset that later benchmarks reuse, describe the first job as "finding needles in a haystack" [1]. Each benchmark below isolates one or more of these jobs, and its score describes that job only.

Expert labels are expensive to make. CUAD's annotators had 70 to 100 hours of contract review training, and three further annotators verified each label [1]. The authors value the labels at over $2 million, assuming each of 9,283 pages was reviewed at least four times at 5 to 10 minutes a page and $500 an hour [1]. MAUD, a merger agreement dataset, took over 10,000 hours of work by law students and lawyers [5].

CUAD is published under a CC BY 4.0 license [2], and later benchmarks reuse its contracts: LegalBench [3], CLAUSE [9], ContractEval [13], ContractScrub [15], LegalBench-RAG [18] and a 2026 re-evaluation [17]. Its contracts come from EDGAR, the SEC's public filing system [15], as do LexGLUE's contract provisions [6]. Most of the evidence reviewed here therefore comes from English-language U.S. commercial contracts filed with the SEC.

Earlier and adjacent work

CLAUDETTE, published in the journal Artificial Intelligence and Law in 2019, detected potentially unfair clauses in 50 online terms of service with an F1 of about 0.81 using classical machine learning [7]. Those are consumer terms, a different setting from negotiated contracts. ACORD, from ACL 2025 Findings, tests clause retrieval for drafting with 114 queries and over 126,000 rated query and clause pairs [8]. LAUKIN, a June 2026 preprint, holds 14,727 clause pairs from 204 Australian, UK and Indian contracts and reports a best macro-F1 of 65.11% for judging whether two clauses are equivalent [19].

The benchmarks also span five years of models. CUAD, ContractNLI and MAUD report fine-tuned encoder models from 2021 to 2023 [1, 4, 5]. LegalBench reports commercial models queried in mid 2023 [3]. ContractEval, CLAUSE and ContractScrub report models from 2025 and 2026 [9, 13, 15]. A gap between two benchmarks can come from the task, the metric or the model generation, and the sources rarely separate the three.

Method

What was examined. Eight benchmarks that test language models on contract text: CUAD [1], LegalBench's contract tasks [3], ContractNLI [4], MAUD [5], ContractEval [13], CLAUSE [9], ContractScrub [15] and VLAIR, an industry report on LLM-based legal tools [21]. Supporting sources cover classification, retrieval and scoring [6, 7, 8, 17, 18, 19, 23, 24], comparisons with lawyers [20, 25, 26] and how accuracy claims are read [10, 11, 12].

What was recorded. For each benchmark: input, required output, matching and scoring rule, best result, year, model generation, review status, data availability and who ran it.

Sources. Peer-reviewed papers come first [1, 3, 4, 5, 6, 7, 8, 9, 10], then NIST [11] and the FTC [12], then workshop papers and preprints [13, 14, 15, 17, 18, 19, 20], industry reports and live leaderboards [21, 22, 23, 24], and trade press and a vendor page [25, 26]. Each source was opened and its passages read on 24 October 2026, and the dated statements were checked against their sources again on 7 November 2026.

How. Where a paper's table and text differ, the table is used and the difference reported. Three kinds of figure are our computation from published tables: LegalBench means and the item total over its 38 CUAD tasks [3] and the share of CUAD test questions with no answer [17]. Model, tool and vendor names appear only in source titles. No new experiments were run, and no product is ranked.

Benchmark	Review status	Data availability, as stated
CUAD	NeurIPS 2021 Datasets and Benchmarks track, peer reviewed [1]	Public, CC BY 4.0 [2]
LegalBench	NeurIPS 2023 Datasets and Benchmarks track, peer reviewed [3]	CUAD tasks under CC BY 4.0 [3]
ContractNLI	Findings of EMNLP 2021, peer reviewed [4]	Released with the paper [4]
MAUD	EMNLP 2023, peer reviewed [5]	Released with the paper [5]
ContractEval	Preprint [13], one-page abstract at the NLLP 2025 workshop [14]	Uses CUAD's test split [13]
CLAUSE	Findings of EACL 2026, peer reviewed [9]	By request form [9]
ContractScrub	ICML AI4Law workshop paper, preprint [15]	Page held only a README on 24 October 2026 [16]
VLAIR	Industry report, not peer reviewed [21]	A sample of questions and answers [21]

Findings

1. CUAD: clause spans scored at fixed recall

CUAD holds 510 contracts of 25 types, from a few pages to over one hundred, with 13,101 labeled clauses in 41 categories [1]. The text for any one label averages about 0.25% of a contract [1]. A model outputs the start and end of each relevant span, and a prediction matches when its word-level Jaccard similarity with the annotation is at least 0.5 [1]. Scores are reported as area under the precision and recall curve (AUPR) and as precision at fixed recall.

Model (2021 paper)	AUPR	Precision at 80% recall	Precision at 90% recall
Base-size encoder, 2018	32.4%	8.2%	0.0%
Large encoder	48.2%	38.1%	0.0%
Extra-large encoder, 2021	47.8%	44.0%	17.8%

Source: [1].

The best model depends on the measure: the extra-large encoder leads at fixed recall, and the large encoder has the higher AUPR [1]. The authors argue that recall matters more because the task is a search, and that 80% recall "may already be reasonable for some lawyers" [1]. At about 30% precision, they note, a lawyer reads about two irrelevant clauses for every relevant one [1]. Results ranged from near 100% AUPR on some categories to about 20% on others [1].

A September 2026 re-evaluation examined what headline scores leave out [17]. CUAD's published test file has 4,182 questions, one per category in each of 102 contracts. Of those, 1,244 carry 2,643 reference spans and 2,938 have no annotated answer [17], 70.3% by our computation. A public fine-tuned checkpoint set to a 90% recall target recovered 82.2% of reference spans at 8.2% precision [17]. It returned text for 1,335 of the 2,938 questions with no answer, a 45.4% false-positive rate with a 95% interval of 43.3% to 47.9%, and only 20.3% of its returned strings matched a reference [17]. Only nine of the 41 categories have at least 30 positive and 30 negative contracts in the test split [17]. The study covers one checkpoint and threshold policy, and the author discloses earlier test access [17].

2. The same clauses as yes-or-no questions (LegalBench)

LegalBench has 162 tasks covering six types of legal reasoning, with results for 20 models [3]. It includes 38 binary tasks built from CUAD, 14 from ContractNLI and 34 from MAUD [3]. For each of 38 CUAD categories, the authors paired clauses from that category with an equal number sampled from other categories, and

scored answers by balanced accuracy [3]. Each item is one clause shown alone. Mean item length runs from 42.4 to 99.7 words across the tasks, which hold 18,208 items, by our computation [3].

Result on the 38 CUAD tasks	Balanced accuracy
Three commercial models, means	92.9%, 89.2% and 88.1%
Top model, range across tasks	77.8% to 100.0%
Five open models of about 13 billion parameters, means	50.5% to 81.4%

Source: our computation from the paper's per-task tables [3]. Chance is 50%.

The paper reports that the three commercial models "all have a balanced-accuracy ≥ 88%" on these tasks [3]. Governing law scored 98.3% to 99.3% for the three, and uncapped liability 70.4% to 85.4% [3]. The authors find that performance degrades on longer texts and multi-class questions: the commercial models averaged 74% to 75% on supply chain disclosures of one to two pages, and the model that led on the CUAD tasks averaged 47.8% on the MAUD tasks [3]. They add three caveats. Results were retrieved between May and August 2023, better prompts may score higher, and the CUAD, ContractNLI and MAUD tasks were already public, so some may have leaked into pretraining data [3].

The live LegalBench leaderboard showed a top score of 88.56% on 13 October 2026, with the first three models within 0.26 points [23]. That score spans LegalBench's six reasoning categories, contract clause interpretation among them [3, 23].

Fixed label sets score high under simple methods too. On LexGLUE's LEDGAR task, 80,000 contract provisions from SEC filings in 100 categories, a TF-IDF baseline reached 87.0 micro-F1 against 88.3 for the best transformer [6]. On its UNFAIR-ToS task, where 95.5% of sentences carry no unfairness label, micro-F1 runs from 94.7 to 96.0 and macro-F1 from 75.0 to 83.0 [6]. A micro average leans on the common class.

3. Entailment with evidence spans (ContractNLI)

ContractNLI gives a system a whole contract and hypotheses such as "Some obligations of Agreement may survive termination of Agreement." The system labels each one entailed, contradicting or not mentioned, and identifies evidence spans [4]. The dataset covers 607 non-disclosure agreements and 17 hypotheses [4]. The authors found that existing models "fail badly" and built a stronger baseline [4]. Their best 2021 model reached 0.892 accuracy against 0.674 for a majority vote, with 0.859 F1 on entailment and 0.405 on contradiction [4]. Contradictions scored worse because of imbalanced labels [4].

The paper isolates negation by exception: a general rule followed by an exception, such as a ban on disclosing confidential information that then permits disclosure to employees [4]. The non-local case, with rule and exception sometimes pages apart, appears in 7% of document and hypothesis pairs and 44% of documents [4]. Without it, accuracy was 0.98 on the majority label and 0.72 on the minority label. With it, the figures were 0.90 and 0.00 [4]. In LegalBench's excerpt-level versions of 14 ContractNLI tasks, which remove the document search, the model that led on its CUAD tasks scored 80.4% to 98.5% [3].

4. Merger agreement deal points (MAUD)

MAUD is based on the American Bar Association's 2021 Public Target Deal Points Study, with over 39,000 examples and 47,000 annotations [5]. Its main dataset holds 20,623 examples of deal point text extracted by experts from 152 merger agreements, and models answer 92 questions, which become 144 once multilabel questions are split into binary ones [5]. The model reads text a lawyer already extracted, so MAUD tests interpretation of a located clause.

The best fine-tuned model reached 57.8% AUPR overall against 16.8% for random answers [5]. By category, the best results ran from 100% on remedies to 50.9% on material adverse effect and 48.2% on conditions to closing [5]. Length is part of the difficulty. Half of all example texts exceed 512 tokens, and every model but one was truncated to that length [5]. The three hardest categories have the longest texts and are the ones attorneys find most difficult to review [5].

A system that retrieves before it reads adds a search step. LegalBench-RAG, a 2024 preprint from a retrieval company, traced LegalBench questions to their source documents [18]. Its merger agreements average 351,476 characters, against 55,827 for CUAD contracts [18]. With a basic pipeline and 64 retrieved chunks, recall was 25.62% for MAUD, 75.71% for CUAD and 86.57% for ContractNLI [18].

5. Clause extraction by current LLMs (ContractEval)

ContractEval returns to CUAD's search task with 2025 models [13]. Each contract is given in full, from 0.6 thousand to 301 thousand characters, and the model returns the clause text or "no related clause" [13]. Labels are 30% positive and 70% negative, and a true positive requires the prediction to fully cover the labeled span [13]. The abstract counts 4 proprietary and 15 open models, and the results table has 19 rows built on 12 distinct models, because reasoning ("thinking") modes and quantized versions are separate rows [13]. The paper gives 4,128 test data points [13], while CUAD's published test file has 4,182 questions [17]. The paper does not explain the difference.

Configuration group (2025)	F1	False "no related clause" rate
Two leading proprietary models	0.644 and 0.641	0.072 and 0.071
Other two proprietary models	0.523 and 0.497	0.025 and 0.011

Configuration group (2025)	F1	False "no related clause" rate
Best open configuration	0.540	0.110
All 19 configurations	0.071 to 0.644	0.000 to 0.306

Source: Table 3 [13].

Averages hide the clause types that matter. The top model approached or exceeded 0.9 F1 on governing law and parties, while both it and the open model compared with it in the paper's category chart scored near zero on uncapped liability, joint IP ownership and notice period to terminate renewal [13]. The authors call these clauses "important and of high risks in real life" [13]. The false rate measures missed clauses directly: one quantized configuration missed relevant clauses in nearly a third of its answers [13].

The paper's text says reasoning mode reduces correctness, with drops in F1 across most models, including the 4B, 8B and 14B sizes of one open model family [13]. Its Table 3 shows drops for the 4B model, from 0.411 to 0.075, and the 14B model, from 0.473 to 0.387, but a rise for the base 8B model, from 0.530 to 0.540 [13]. Reasoning cut that model's false rate from 0.248 to 0.110 [13]. The abstract says most models perform "at a level comparable to junior legal assistants" [13]. The paper reports no human baseline, so it does not measure that comparison.

6. Planted discrepancies (CLAUSE)

CLAUSE, in the Findings of EACL 2026, perturbs real contracts to test whether models detect and explain discrepancies [9]. It holds 4,711 CUAD and 2,803 ContractNLI contracts with 12,869 and 11,086 perturbations in ten anomaly categories [9]. The authors call it "the first fully AI-generated legal contradiction dataset" [9]. One language model generated the perturbations, and other models judged the explanations [9].

In the binary detection task, models were "provided with both the original and perturbed documents" [9]. On CUAD categories, the top model reached 62.8% to 65.4% F1, with recall of 90.1% to 95.8% and accuracy of 47.1% to 49.8% [9]. Across all 80 cells of the table, accuracy ran from 9.2% to 55.8% [9]. Explanations scored "high Clarity (often 4.0+) but low Completeness (<2.0)" on a 0 to 5 scale, and matching legal citations stayed below 14% [9]. One legal expert and three NLP researchers validated the data voluntarily, and access is by request form [9].

7. Final review for errors and inconsistencies (ContractScrub)

ContractScrub defines contract scrubbing as "the final review of transactional agreements for errors and inconsistencies" [15]. It is an August 2026 ICML AI4Law workshop paper by researchers at a legal-technology company and Imperial College London, testing general models [15]. It holds 3,014 annotated tasks in 44 CUAD contracts of about 10 to 15 pages [15]. Nine lawyers, each with at least eight years of practice,

recorded defined terms and existing errors, then inserted further errors across eight categories, with defined-term extraction as a ninth [15]. Recall is the primary metric because "it is much easier to check whether a flagged issue is real than to identify issues that were not previously known" [15].

Result (August 2026)	Recall	Precision	F1	Cost per contract	Time
Top configuration by recall	0.750	0.580	0.632	$1.38	533 s
Top configuration by F1	0.744	0.616	0.655	$0.19	76 s
Lowest F1 with reasoning	0.438	0.268	0.316	$0.04	194 s
Top without reasoning	0.643	0.463	0.525	$0.42	22 s

Source: Table 2 [15].

The paper is inconsistent in two places. Its introduction states that "all F1 scores are below 0.650", while Table 2 shows 0.655 and the results section says "nearly all" [15]. It says 9 models were evaluated, and Table 2 names ten [15].

Categories separate surface checks from reading. Mean recall across models was .835 for defined terms and .781 for unused defined terms, against .362 for incorrect party references, .427 for incorrect capitalization in context and .351 for undefined capitalized terms [15]. Recall on wrong section references fell as the distance between the reference and the section it points to grew, most past ten thousand characters, about five to six pages [15]. The authors judge that full automation likely needs more than "the 75% recall currently attained" [15]. They note the 44-contract scale and a required JSON output that may depress scores [15]. They state that they release the benchmark, and on 24 October 2026 its dataset page held only a README file [15, 16].

8. Tools against a lawyer baseline (VLAIR)

VLAIR, updated 27 February 2025, evaluated four LLM-based legal tools on seven tasks against a lawyer control group [21]. Participating law firms supplied the questions and reference answers [21]. The lawyers, sourced through an alternative legal service provider, did not know they were in a study and had two weeks to respond [21]. Some questions came with documents of 100 to 900 pages [21]. A language model graded each check against one element of a reference response, and failing scores got a second review [21]. The report states that this grading "is not sufficient to replace an expert human grader" and discloses that its publisher has a customer relationship with one or more of the participants [21].

Task (February 2025)	Lawyer baseline	Top tool	Tools the report marks above the baseline
Data extraction	71.1 ± 3.2	75.1 ± 3.2	2 of 4
Document Q&A	70.1 ± 5.2	94.8 ± 2.5	3 of 4
Redlining	79.7 ± 4.8	65.0 ± 5.0	0 of 2 tested

Source: results table [21].

Data extraction had 30 questions over 29 documents and 204 checks, and a good answer needed the verbatim text and its source location [21]. On the hardest question, 40 fields across three credit agreements, the top tool extracted 61.7% of fields correctly and the lawyers 39.2% [21]. Asked for warranty disclaimers, with three clauses in the expected answer, every tool found at least one and none found all three [21]. One tool's redlining score reads 65.0 in the results table and 59.4% in the text [21]. The tools were 6 to 80 times faster than the lawyers [21]. VLAIR's October 2025 follow-up covered legal research questions [22].

Two earlier studies, both run by vendors, also compare review tools with lawyers on contracts.

Study	Material	Lawyers	Top tool result	Run by
Vendor study, 2024	10 procurement contracts	F 0.860 junior lawyers, F 0.874 outsourced reviewers	F 0.871	Legal-technology company staff, preprint [20]
NDA study, 2018	5 NDAs, 20 lawyers	85%	94%	The vendor tested [25, 26]

In the 2024 vendor study, senior lawyers set the ground truth by majority consensus and agreed with each other least of any group, at a Cronbach's alpha of 0.719 [20]. Every group scored lower on locating issues than on determining them, and outsourced reviewers led on location at 0.77 against 0.74 for the top model [20]. Senior lawyers took 43.46 minutes and $75.92 per document, and the models 0.73 to 4.70 minutes and $0.02 to $1.24 [20]. The 2018 study used five public NDAs and 30 issue types set by its organizers [26]. Lawyers averaged 92 minutes for all five, against 26 seconds for the system [25].

9. Why task framing moves the scores

The findings share six mechanisms.

Framing choice	Example	Effect on the reported number
The clause is given, or must be found	LegalBench items against CUAD and ContractEval [1, 3, 13]	Classifying a given clause scores far higher
Class balance	Balanced LegalBench tasks against 70% absent clauses in CUAD's test [3, 17]	A balanced set hides answers given when nothing is there
Denominator	82.2% of references found, 8.2% precision, 20.3% of returned strings matching [17]	One run yields very different figures
Matching rule	Jaccard of 0.5 in CUAD, full coverage in ContractEval, corrected matching in the re-evaluation [1, 13, 17]	The rule changes the reported score
Unit of scoring	27.08% of tasks against 96.30% of criteria [24]	Whole-task scoring is far lower
Recall and precision trade	CLAUSE F1 near 63% to 65% with accuracy below 50% [9]	A high-recall strategy lifts F1

The scoring unit shows the effect plainly. On a legal agent benchmark released by a legal technology company and scored by an evaluation company, the leading model passed 27.08% of 120 held-out tasks on 5 November 2026, while 96.30% of its individual criteria passed, because a task passes only when every criterion does [24]. ACORD shows a similar gap in ranking: its best method reached 79.1% NDCG@5 while its normalized five-star precision at five was 17.2% [8].

Length and distance add pressure under every framing. ContractScrub recall fell past ten thousand characters between a section reference and its target [15], half of MAUD's example texts exceed the 512 tokens that all but one model could read [5], ContractNLI's minority label failed when an exception sat far from its rule [4], and retrieval recall on merger agreements was 25.62% at 64 chunks [18].

10. What the spread means for a buyer's own test

NIST says accuracy measurements "should always be paired with clearly defined and realistic test sets" that represent expected use, should consider false positive and false negative rates, and may be broken down by data segment [11]. Its MEASURE 2.3 asks that performance be shown "for conditions similar to deployment setting(s)" [11]. Applied to contracts, the findings suggest these elements:

1. **The buyer's own documents.** Most public benchmarks here use SEC-filed U.S. contracts [6, 15] that may sit in training data [3]. The one study of Australian, UK and Indian contracts reports a best macro-F1 of 65.11% on its task [19]. A held-out set, like the legal agent benchmark's held-out test set [24], lowers the risk

that test items were in training data.

2. **Absent clauses at their real rate,** with a count of answers given when no clause exists [17].
3. **Recall by clause type,** with enough examples per type. Averages hid near-zero categories in ContractEval [13], and only nine of CUAD's 41 categories meet a 30-and-30 support floor [17].
4. **A matching rule fixed in advance,** since the rule changes the score [1, 13].
5. **Location as well as answer.** Every group in the 2024 vendor study scored lower on locating issues [20], and VLAIR required source locations [21].
6. **Long documents and distant references,** where scores fell in ContractScrub and ContractNLI [4, 15].
7. **Repeat runs.** VLAIR reports margins of about ±3 points on extraction [21], so gaps of a few points carry little weight.
8. **Cost and time.** With reasoning on, ContractScrub configurations cost $0.04 to $1.52 and took 23 to 533 seconds per contract [15].

Claims made without such tests have drawn scrutiny. In February 2025 the FTC finalized an order against an online legal service that had promoted itself as "the world's first robot lawyer" [12]. The complaint alleged that the company "did not test whether its 'AI lawyer' operated to the level of a human lawyer", and the order bars it from claiming lawyer-level performance without sufficient evidence and requires $193,000 in monetary relief [12]. A preregistered study of legal research tools, a different task, found providers' claims to eliminate or avoid hallucinations overstated, with rates of 17% to 33% [10].

Implications

For legal operations leads and in-house counsel. A headline accuracy figure describes one task under one scoring rule. The sources show yes-or-no clause classification at 88.1% to 92.9% in 2023 [3], whole-contract extraction at 0.644 F1 at best [13] and final error review at 0.750 recall at best [15]. The lowest-scoring categories, such as uncapped liability, are ones the ContractEval authors call high risk [13], and the ContractScrub authors judge that full automation needs more recall than any model reached [15].

For procurement teams comparing tools. The sources point to asking for the task definition, the matching rule, the false-answer rate on absent clauses, recall by clause type and the documents used [11, 13, 17]. Every tool-against-lawyer study found here was run by a vendor or by an evaluator with a disclosed customer relationship to participants [20, 21, 25]. A test on the buyer's own documents, scored by the buyer's reviewers, answers what those studies cannot.

For researchers and benchmark builders. No source runs the same model on the same clauses as both a yes-or-no question and an extraction task. LLM-as-judge grading in VLAIR and CLAUSE has no published validation against lawyer grading: VLAIR gives failing scores a second review, and CLAUSE compares its judge with NLP experts' scores on a sample of outputs [9, 21]. Data release decides who can reproduce a result: CUAD is public [2], CLAUSE is by request [9] and ContractScrub's dataset page held no data on 24 October 2026 [16].

Rillor's document and regulation intelligence work uses LLMs to extract parties, defined terms, obligations, conditions and deadlines from contracts as fields, each linked to the exact passage it came from and accepted by a reviewer before the record is used. Reviewer corrections go into the test set for the next extraction run.

Limits of this analysis

- This is a synthesis of published results. It runs no new tests and reports figures as the sources give them.
- The benchmarks differ in metric, matching rule, model generation and year, so their scores cannot be ranked against one another. The framing effect is inferred across papers.
- Several sources are workshop papers or preprints [13, 14, 15, 17, 18, 19, 20], and several are written by vendors or evaluators with commercial ties [15, 18, 20, 21, 24, 25, 26].
- Figures marked as our computation come from published tables [3, 17]. Leaderboard figures are as read on 7 November 2026 and change over time [23, 24].
- No source measures training-data contamination for these datasets, although LegalBench flags the risk [3].
- Coverage is mostly English-language U.S. contracts from SEC filings. No benchmark here covers scanned contracts, OCR errors or amendments spread across documents. LAUKIN is the only source on other common-law jurisdictions [19].
- No source tests a commercial contract review product on a public dataset. VLAIR tests products on private questions [21]. No 2026 study of tools against lawyers on contract tasks was found.
- The 2018 NDA study's full report was not reachable, so its description rests on trade press and the vendor's page [25, 26].
- This report is not legal advice and ranks no product.

Sources

1. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. Dan Hendrycks, Collin Burns, Anya Chen and Spencer Ball, NeurIPS 2021 Track on Datasets and Benchmarks, arXiv 2103.06268, version 2, 8 November 2021. https://arxiv.org/abs/2103.06268

2. Contract Understanding Atticus Dataset (CUAD). The Atticus Project, viewed 24 October 2026. https://www.atticusprojectai.org/cuad

3. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré and others, NeurIPS 2023 Datasets and Benchmarks Track, arXiv 2308.11462, 20 August 2023. https://arxiv.org/abs/2308.11462

4. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. Yuta Koreeda and Christopher D. Manning, Findings of the Association for Computational Linguistics: EMNLP 2021, arXiv 2110.01799, 5 October 2021. https://arxiv.org/abs/2110.01799

5. MAUD: An Expert-Annotated Legal NLP Dataset for Merger Agreement Understanding. Steven H. Wang, Antoine Scardigli, Leonard Tang and others, Proceedings of EMNLP 2023, arXiv 2301.00876, version 3, 24 November 2023. https://arxiv.org/abs/2301.00876

6. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. Ilias Chalkidis and others, ACL 2022, arXiv 2110.00976, version 4, 8 November 2022. https://arxiv.org/abs/2110.00976

7. CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service. Marco Lippi, Przemyslaw Palka, Giuseppe Contissa and others, Artificial Intelligence and Law 27 (2019), arXiv 1805.01217, version 2, 18 February 2019. https://arxiv.org/abs/1805.01217

8. ACORD: An Expert-Annotated Retrieval Dataset for Legal Contract Drafting. Steven H. Wang and others, Findings of ACL 2025, arXiv 2501.06582, 11 January 2025. https://arxiv.org/abs/2501.06582

9. Better Call CLAUSE: A Discrepancy Benchmark for Auditing LLMs Legal Reasoning Capabilities. Manan Roy Choudhury, Adithya Chandramouli, Mannan Anand and Vivek Gupta, Findings of the Association for Computational Linguistics: EACL 2026, pages 5776 to 5818, March 2026. https://aclanthology.org/2026.findings-eacl.305/

10. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning and Daniel E. Ho, Journal of Empirical Legal Studies 22(2), 216 to 242, 23 April 2025, arXiv 2405.20362. https://arxiv.org/abs/2405.20362

11. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. National Institute of Standards and Technology, January 2023. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

12. FTC Finalizes Order with DoNotPay That Prohibits Deceptive 'AI Lawyer' Claims, Imposes Monetary Relief, and Requires Notice to Past Subscribers. U.S. Federal Trade Commission, 11 February 2025. https://www.ftc.gov/news-events/news/press-releases/2025/02/ftc-finalizes-order-donotpay-prohibits-deceptive-ai-lawyer-claims-imposes-monetary-relief-requires

13. ContractEval: Benchmarking LLMs for Clause-Level Legal Risk Identification in Commercial Contracts. Shuang Liu, Zelong Li, Ruoyun Ma, Haiyan Zhao and Mengnan Du, arXiv 2508.03080, 5 August 2025. https://arxiv.org/abs/2508.03080

14. ContractEval: Benchmarking LLMs for Clause-Level Legal Risk Identification in Commercial Contracts. Shuang Liu, Zelong Li, Ruoyun Ma, Haiyan Zhao and Mengnan Du, Proceedings of the Natural Legal Language Processing Workshop 2025, Association for Computational Linguistics, November 2025, page 291. https://aclanthology.org/2025.nllp-1.19/

15. ContractScrub: A benchmark for final review of legal contracts. Yejin Bang, Kirsty Fielding, Brandan Oliver, Brian Birke, Nabeel Seedat and Andrew M. Bean, ICML AI4Law Workshop, arXiv 2608.20204, 20 August 2026. https://arxiv.org/abs/2608.20204

16. tri-fair-lab/contract_scrub, dataset page. Hugging Face, viewed 24 October 2026. https://huggingface.co/datasets/tri-fair-lab/contract_scrub

17. Abstention Errors and Segment Support in CUAD: An Auditable Contract-Extraction Case Study. Zeki Emre Tekin, arXiv 2609.13274, 7 September 2026. https://arxiv.org/abs/2609.13274

18. LegalBench-RAG: A Benchmark for Retrieval-Augmented Generation in the Legal Domain. Nicholas Pipitone and Ghita Houir Alami, arXiv 2408.10343, 19 August 2024. https://arxiv.org/abs/2408.10343

19. LAUKIN: A Multi-jurisdictional Common Law Contract Dataset. Amrita Singh, Aditya Joshi, Jiaojiao Jiang, Hye-young Paik and May Fong Cheong, arXiv 2606.13184, 11 June 2026. https://arxiv.org/abs/2606.13184

20. Better Call GPT, Comparing Large Language Models Against Lawyers. Lauren Martin, Nick Whitehouse, Stephanie Yiu, Lizzie Catterson and Rivindu Perera, arXiv 2401.16212, 24 January 2024. https://arxiv.org/abs/2401.16212

21. Vals Legal AI Report (VLAIR). Vals AI with Legaltech Hub, updated 27 February 2025, viewed 24 October 2026. https://www.vals.ai/vlair

22. VLAIR, Legal Research. Vals AI, updated 14 October 2025, viewed 24 October 2026. https://www.vals.ai/industry-reports/vlair-10-14-25

23. LegalBench, benchmark results. Vals AI, updated 13 October 2026, viewed 7 November 2026. https://www.vals.ai/benchmarks/legal_bench

24. HLAB-Vals-1.1, benchmark results. Vals AI, updated 5 November 2026, viewed 7 November 2026. https://www.vals.ai/benchmarks/hlab

25. LawGeex Hits 94% Accuracy in NDA Review vs 85% for Human Lawyers. Artificial Lawyer, 26 February 2018. https://www.artificiallawyer.com/2018/02/26/lawgeex-hits-94-accuracy-in-nda-review-vs-85-for-human-lawyers/

26. AI vs Lawyers: The Ultimate Showdown. Ilana Waxman, Superlegal, 24 October 2022, viewed 24 October 2026. https://www.superlegal.ai/blog/aivslawyer/

About this report

Published by Rillor on 7 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. What benchmarks show about LLM contract review accuracy. 7 November 2026. https://rillor.com/insights/llm-contract-review-accuracy-benchmarks

Notices

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Send a sample of the rulebooks, regulations, contracts or filings you work from and the fields you need from them. Contact us at rillor.com/contact?type=document-intelligence or contact@rillor.com. More on document and regulation intelligence: rillor.com/si/document-intelligence