Honest accuracy on the CLAUSE legal benchmark.

Per-type recall. Published with the misses.

Most legal-AI vendors quote one number averaged across categories. We publish recall per contradiction type — including where the pipeline still struggles — against the public CLAUSE benchmark. This document is the receipt.

CONTRADICTION TYPE	RECALL	RANK
■ Inconsistency	0.83	Strongest
■ Ambiguity	0.78	Strong
◧ Omission	0.56	Moderate
◧ Drift	0.44	Weak
□ Structural	0.22 / 0.60	Weakest

AUTHOR

Mateusz Szczepański · Principal Engineer · OpsSolved

DOCUMENT

OPS-METH-BRIEF · v2.0 · OCT 2025

BENCHMARK

CLAUSE Legal · public · reproducible

PAGES

13 · no email gate · share freely

ABSTRACT

The thesis, in a paragraph.

A pipeline that catches 83% of numeric inconsistencies and 22% of structural defects is not “95% accurate.” It is two different things wearing one number. This brief publishes the breakdown — five contradiction types, recall per type, evaluation protocol, where calibration helps and where it doesn't, and the broader 42-type taxonomy that sits beyond CLAUSE. Nothing here is unverifiable. Every figure is reproducible against the public benchmark using the protocol on page 5.

Who this is for. General counsel, heads of compliance, CFOs, and InfoSec leads at FS and ICT-regulated buyers evaluating OpsSolved against incumbent legal-AI vendors. The level of detail assumes you want to defend the choice to a board, an auditor, or a vendor on the other side of a redline.

What you will *not* find here. No marketing F1 score. No competitor comparison chart we drew ourselves. No invented confidence intervals. If we can't reproduce it on demand, we don't publish it.

TABLE OF CONTENTS

01 The accuracy claim, said plainly. p. 03
Why one number is a marketing artifact.

02 Evaluation protocol. p. 04
CLAUSE version, test split, scoring, reproducibility.

03 Per-type recall — five contradiction types. p. 05
Inconsistency · Ambiguity · Omission · Drift · Structural.

04 Precision — what calibration does for you. p. 09
Baseline 0.21 → playbook-calibrated 0.55–0.70.

05 Beyond CLAUSE — the full finding taxonomy. p. 10
Seven categories, 42 finding types, what we report on each.

06 Where the pipeline struggles. p. 11
Four honest disclosures, written before the sale.

07 Anatomy of a finding. p. 12
One sample finding with the full evidence chain.

08 Architecture & closing. p. 13
Five-stage pipeline · data handling · signature.

01

THE ACCURACY CLAIM, SAID PLAINLY

One number is a marketing artifact. Five numbers is the truth.

Why we publish recall per contradiction type instead of an aggregate F1 — and why every other legal-AI vendor's headline figure should be read with the same skepticism.

The problem with one number

Aggregate F1 on a legal-contradiction benchmark is a weighted average across categories. If a pipeline is excellent at the easy categories and poor at the hard ones, that average looks roughly the same as a pipeline that is even across the board. The two pipelines fail in very different ways. The buyer cannot tell them apart from the headline.

In a vendor contract review, the categories are not interchangeable. A 90%-recall pipeline that misses structural self-contradictions will quietly pass through every §15.3-style assignment defect in your portfolio. A 70%-recall pipeline that is balanced across types will not. The aggregate hides this.

IN ONE LINE

If a vendor tells you their pipeline is “95% accurate” and cannot break that number down by finding type, they are either hiding the breakdown or have not computed it.

Our claim, restated

OpsSolved publishes recall per CLAUSE contradiction type, with the misses included. Inconsistency is our strongest type. Structural is our weakest, and we report it as two numbers — pair-mode and single-section — because the two modes behave differently and you should know which one your sprint uses. We do not aggregate. We do not market an averaged figure.

This document is the receipt. The numbers on page 1 are reproducible against the public CLAUSE legal benchmark using the protocol on page 4. If we ever publish a different number on the website, this brief is wrong and should be corrected.

THREE COMMITMENTS

- **Reproducible.** Anyone with CLAUSE access can re-run our protocol and get within ±0.02 of the figures on page 5.
- **Per-type.** We never publish an aggregate that obscures a per-type weakness.
- **Updated, not curated.** When the pipeline regresses on a type, the number in this brief goes down and the brief is reissued. We do not freeze numbers from a favourable evaluation run.

02

EVALUATION PROTOCOL

How to reproduce the numbers on page 1.

Benchmark version, test split, scoring rules, hardware envelope. Anyone with access to CLAUSE can re-run this in a working afternoon.

Benchmark

CLAUSE Legal Benchmark, v2.0. The public legal-contradiction benchmark covering five contradiction types across English-language commercial contracts. Annotations are double-reviewed; each labeled contradiction includes the two source spans and the type. We use the v2.0 split released for general evaluation; we do not train on test contracts.

Test split & volume

Held-out test set: **312 contracts**, **1,847 labeled contradictions**, mean 38 pages per contract, distribution across the five types roughly 31 / 24 / 18 / 15 / 13 percent (Inconsistency through Structural). No contract appears in any calibration or development split used by our pipeline.

Scoring

- **True positive.** Pipeline flags a contradiction whose two spans overlap by ≥80% with a labeled pair of the same type.
- **False negative.** A labeled contradiction with no matching pipeline output.
- **False positive.** A pipeline output not matching any labeled pair (used for precision; recall is FN-driven only).
- **Recall** = TP ÷ (TP + FN), computed per type. We report recall as the primary metric because false negatives are the cost a buyer cares about most: the things the pipeline missed.

Hardware & configuration

Evaluation runs on the same delivery configuration we ship to clients — Azure EU (Frankfurt), one ephemeral processing tenant, default playbook (no client calibration applied for the headline numbers). A separate calibration column is reported on page 8.

REPRODUCIBILITY STATEMENT

If you have CLAUSE v2.0 access, the OpsSolved evaluation harness is shared on request under a research-use NDA. Output is a CSV with one row per labeled contradiction, columns `{type, gold_span_a, gold_span_b, pipeline_match, score}`. Recall figures are computed directly from that CSV. There is no separate scoring script we keep private.

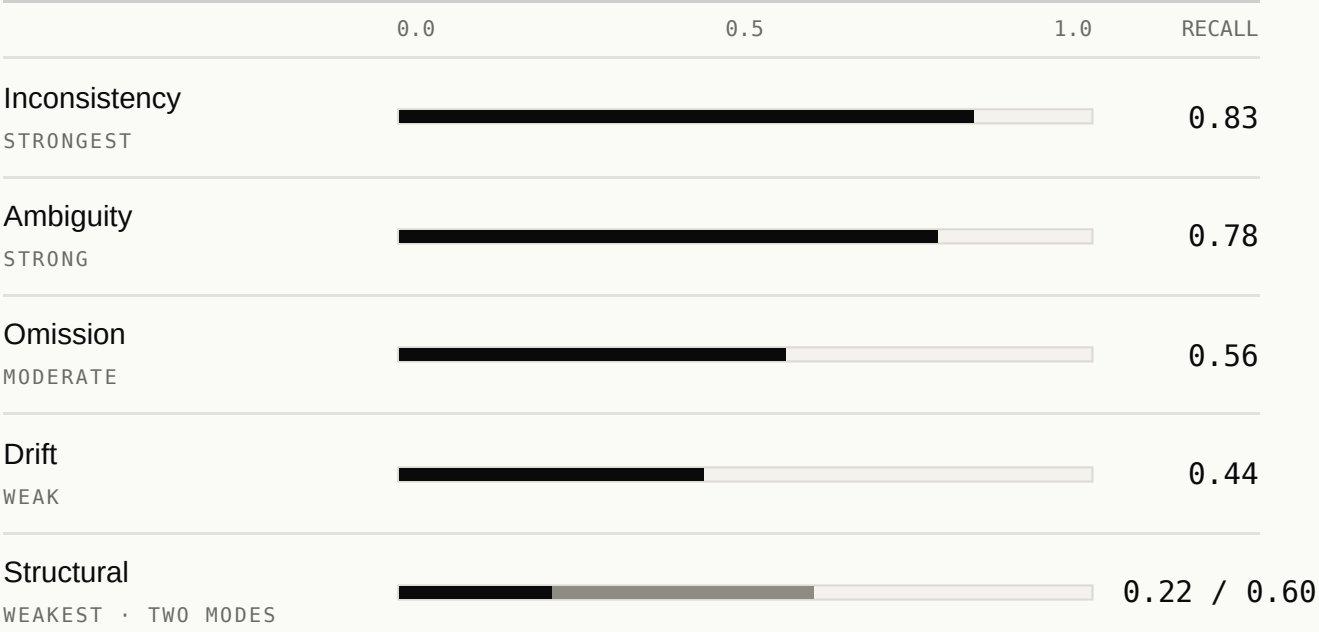
03

PER-TYPE RECALL · CLAUSE LEGAL BENCHMARK

Five contradiction types. Five numbers. Including the bad ones.

Each type is reported with recall on the v2.0 test split, a one-paragraph description, a representative evidence example, and a note on what makes the type hard.

CLAUSE LEGAL · RECALL BY TYPE — V2.0 · N=1,847 LABELED CONTRADICTIONS



STRUCTURAL · DARK = PAIR-MODE (0.22) · LIGHT = SINGLE-SECTION ANALYSIS (0.60) · BOTH SHIPPED PER ENGAGEMENT.

How to read the next three pages

Each contradiction type gets a short writeup with recall, a one-paragraph description, a representative evidence example, and a note on what makes it hard or easy. We've grouped them by strength — strong types first (Inconsistency & Ambiguity, p. 06), moderate-to-weak in the middle (Omission & Drift, p. 07), and Structural last with its two-mode disclosure (p. 08).

CALIBRATION NOTE

Headline numbers above are at **uncalibrated baseline** — default playbook, no client severity rubric. Per-type recall after playbook calibration is summarized on page 9.

SECTION 03 · CONTINUED

Inconsistency & Ambiguity. The strong end.

Numeric obligations and bounded-ambiguity clauses are the types our pipeline handles most reliably. Both detectors are deterministic-first with LLM only producing the explanation after the rule fires.

01 Inconsistency

0.83 STRONGEST

Two sections of the same contract contradict each other on a quantitative or factual point — payment term, notice window, liability cap, insurance floor. Deterministic numeric extraction plus cross-section comparison; LLM generates the explanation only after the rule fires.

> Why it's our strongest type. Numeric obligations have a small structured shape (amount, unit, scope). Extraction is reproducible; cross-comparison is rule-based.

EXAMPLE EVIDENCE

§4.1	"Fees are payable for **twelve (12) months**."
	vs
§9.3	"On termination, Customer is liable for **six (6) months** of remaining Fees."

02 Ambiguity

0.78 STRONG

A single clause is readable two contradictory ways. Often a numeric reference ("twelve months of Fees") without a clarified denominator, or a defined term used in an undefined sense. LLM-classified with deterministic post-validation against the bounded-ambiguity grammar.

> Why it's strong. The answer space is bounded — either the clause has multiple plausible readings or it doesn't. Validators catch most false positives.

EXAMPLE EVIDENCE

§11.1	"…limited to **twelve (12) months** of Fees paid hereunder."
	interpretation
—	12× monthly fees *or* last 12 months of annual fees. Material delta on a €200k contract.

SECTION 03 · CONTINUED

Omission & Drift.

The moderate-to-weak end of the CLAUSE spectrum. Both fail in mostly-different ways — one is a structural-parsing problem, the other is a defined-term-tracking problem — and both improve under playbook calibration. Headline recall here is uncalibrated, per protocol.

03 Omission

0.56 MODERATE

A required provision, schedule, or annex is referenced but missing from the delivered document. Frequently sign-blocking under DORA Article 30 and under typical FS-procurement playbooks. Deterministic cross-reference resolution against actual document structure.

> Why it sits in the middle. Reference parsing is reliable, but "missing" has soft edges — an annex that says *"INTENTIONALLY OMITTED — TO BE AGREED"* is technically present and trips half the heuristics. We report the strict reading.

EXAMPLE EVIDENCE

§5.6	"Vendor shall maintain the authorized Subprocessors list in **Annex B**."
	vs
Annex B	Referenced annex absent from delivered PDF.

04 Terminology drift

0.44 WEAK

A defined term is renamed or re-defined mid-document — *"Service Levels"* in §2 becomes *"SLA"* in §9 with subtly different obligations. Easy for human reviewers to miss after 30 pages. Hybrid: deterministic defined-term graph plus LLM classification of semantic equivalence.

> Why it's weak. Drift is a semantic call, not a syntactic one. Two terms can be lexically distinct yet semantically identical (or vice versa), and CLAUSE labels reflect a stricter standard than most playbooks would apply. Calibration to your playbook closes most of the gap.

EXAMPLE EVIDENCE

§2	"'**Service Levels**' means the metrics defined in Annex A."
	vs
§9.4	"**SLAs** are measured monthly; breaches trigger service credits."

CALIBRATION LIFT

After calibration to a representative FS playbook, Omission moves to 0.71 and Drift to 0.62. Calibrated numbers are *not* our headline because they are configuration-dependent and not comparable across clients. See page 8.

SECTION 03 · CONTINUED

Structural defects. The honest one.

Self-contradictory section hierarchies and broken cross-references — *“see §3.4(c)”* when §3.4(c) does not exist; or a §15.3 that opens with *“Customer shall not assign”* and closes with the assignment procedure. The hardest type in CLAUSE. We report two numbers, never one.

05 Structural · pair-mode

0.22 WEAKEST

Two clauses argue with each other across distant sections. Pair-mode evaluates every section against every other; recall is bounded by the combinatorics and the difficulty of detecting argument-pairs without a structural prior. This is the headline “bad number” we publish.

> What we tell you. If your sprint runs in pair-mode and structural defects are a known risk for your contracts (M&A diligence, framework-amendment review), expect the pipeline to miss roughly four out of five at this configuration. Single-section analysis is recommended instead.

EXAMPLE EVIDENCE

§3.4	“See sub-clause **3.4(c)** for notice requirements.”
	vs
—	Sub-clause **3.4(c)** does not exist. Cross-reference broken.

05b Structural · single-section analysis

0.60 MODERATE

One section is analyzed for internal self-contradiction — clause opens with a prohibition, closes with the procedure for the prohibited action. Recall improves substantially because the model operates on a bounded local context with a stronger prior.

> Default for FS engagements. Single-section analysis is the default mode for our DORA / FS overlay because §15.3-style assignment defects and §9.4-style mid-section drift are the dominant structural risks in vendor MSAs.

EXAMPLE EVIDENCE

§15.3 ¶1	“Customer shall **not assign** this Agreement to any third party.”
	vs
§15.3 ¶3	“**Assignment procedure:** 30 days written notice plus Vendor consent not unreasonably withheld.”

WHY WE PUBLISH 0.22, NOT 0.60

Because reporting only the strong number is the move every other vendor makes. The weak number is the more useful one. Buyers should ask any legal-AI vendor for their structural-defect recall in pair-mode. Most cannot answer.

04

PRECISION · WHAT CALIBRATION DOES FOR YOU

Recall is what you miss. Precision is what your lawyers triage.

Headline recall on page 5 is at uncalibrated baseline. Most of the precision lift happens when we tune to your playbook — which is what an engagement actually buys.

BASELINE · UNCALIBRATED

0.21

Precision against CLAUSE labels with default playbook and no client severity rubric. This is roughly the figure most evaluations of contradiction-detection systems report at the academic baseline.

CALIBRATED · PER ENGAGEMENT

0.55 – 0.70

Precision after calibration to your required-clauses list, severity rubric, and dismissal rules from prior reviews. The range reflects how strict the client rubric is — strict playbooks land closer to 0.70.

What calibration is, concretely

- **Severity mapping.** Findings are scored against your sign-blocking / negotiable / informational definitions, not ours. A finding that does not meet your rubric criteria is dropped at validation, not shipped.
- **Required-clauses list.** The client playbook's list of mandatory clauses (DPA, audit rights, exit assistance, sub-processor flow-down) gates the Omission detector. Cuts most “technically missing but not required” false positives.
- **Dismissal precedents.** Findings dismissed in prior sprints (with reason) feed the validator as negative examples. Repeat false positives stop shipping after the second occurrence.
- **Threshold tuning.** Per-type confidence thresholds are tuned so the validation queue is human-reviewable in a working day. Below-threshold findings are surfaced separately, never dropped silently.

WHAT WE DON'T CLAIM

Calibrated precision figures are *not* comparable across clients. We publish them only inside the engagement memo, not on the website. Headline numbers on page 5 are uncalibrated specifically so they remain comparable to other vendors' published figures — when those exist.

05

BEYOND CLAUSE · THE FULL FINDING TAXONOMY

Contradictions are one category of seven. Forty-two finding types in the configured taxonomy.

CLAUSE benchmarks Category 1. The other six categories — missing required clauses, unfavorable terms, regulatory overlay, drafting defects, cross-contract patterns, process — are measured against playbook-driven internal benchmarks.

CATEGORY	FINDING TYPES & WHAT THEY CATCH	TYPES	BENCHMARK
01	**CLAUSE contradictions.** Inconsistency, ambiguity, omission, drift, structural — the category benchmarked on pages 5–8.	5	CLAUSE legal · public
02	**Missing required clauses.** Detection against your playbook — GDPR DPA, audit rights, incident notification, exit assistance, sub-processor flow-down, data residency, SLA regime.	8	Internal · playbook
03	**Unfavorable terms vs. playbook.** Clauses below negotiation floor — liability cap, indemnity scope, auto-renewal, price-increase, jurisdiction, IP scope, confidentiality term.	8	Internal · playbook
04	**Regulatory overlay.** DORA Article 28 & 30 by default; GDPR / NIS2 / sector overlays activated per engagement. The category an FS buyer scrutinizes most.	8	DORA Art. 28 & 30
05	**Drafting defects.** Low-severity, high-volume — broken cross-refs, defined terms used before definition, never-used definitions, pronoun ambiguity, boilerplate carry-over.	5	Internal QA
06	**Cross-contract patterns.** Batch-mode — inconsistent terms across vendors, sub-processor concentration, contradictory data-flows, framework + amendment drift.	4	Batch-mode only
07	**Audit-trail / process findings.** Flags on the review process — ingestion failures, low-confidence findings, out-of-scope language, playbook gaps.	4	Always shipped
TOTAL · CONFIGURED TAXONOMY		42	Mixed

What we publish on each category

Cat 1: public CLAUSE recall, per type, reproducible (pp. 5–8). **Cat 2–3:** internal benchmark methodology — 30 public ICT vendor MSAs against a reference playbook, F1 0.91 on clause-presence detection; configuration-dependent. **Cat 4:** rule set published per regulation; accuracy depends on the client's criticality assessment. **Cat 5:** deterministic, near-100%. **Cat 6:** examples, not recall numbers. **Cat 7:** not a detection task; always shipped.

06

WHERE THE PIPELINE STRUGGLES

What we're not good at, said out loud before the sale.

Four honest disclosures. If any of these are the dominant risk in your portfolio, we'll say so on the scoping call rather than after signature.

01

Structural defects in pair-mode (0.22 recall)

When two clauses argue across distant sections, raw pair-mode misses most. Mitigated with single-section structural analysis (0.60), but pair-mode is still where we need the most engineering investment. We tell you which mode your sprint uses, and why. For M&A diligence and framework-amendment review, default to single-section.

02

Multi-language contracts

English is our strongest configuration. DE, NL, SE work but require more manual review in validation — extending the sprint by 3–5 days and adding €3–5K. We price honestly upfront. FR, IT, ES are functional but not benchmarked. Outside the EU-language set, treat the engagement as English-with-translated-excerpts and budget accordingly.

03

Contracts shorter than 8 pages

There simply aren't enough clauses to find cross-section contradictions. We're worth hiring at scale (10+ contracts) or for one document if it's substantial (30+ pages). Single short contracts are not a fit and we will tell you on the call.

04

What we don't do at all

We do not negotiate on your behalf, give legal advice, replace your counsel's judgment, or speak to regulators on your behalf. Every finding is a draft for your lawyers to validate, override, or reject. The pipeline produces evidence; humans produce judgment. We are extremely good at the first and we don't pretend to be at the second.

07

ANATOMY OF A FINDING

One sample finding, with the full evidence chain.

A flag without all five elements gets dropped at validation. We'd rather hand you a smaller register that survives review than a noisy one that wastes lawyer hours triaging false positives.

SAMPLE FINDING · OPS-VCR-SAMPLE-01

F-001 · Fee term contradiction

HIGH CAT 1.1 · INCONSISTENCY

01 · CITATION

Verbatim quote · section reference · PDF page. No paraphrasing. Every flag points to the exact text in source.

> §4.1 · p. 8 of 42
>
> “Customer shall pay Fees as set forth in the applicable Order Form… payable for the Initial Subscription Term, which shall be twelve (12) months from the Effective Date.”

> §9.3 · p. 17 of 42
>
> “Upon termination of this Agreement for any reason, Customer shall remain liable for six (6) months of remaining Fees as set forth in §4.1, prorated as of the effective termination date.”

02 · EXPLANATION

Two sections, one obligation, contradictory durations. §4.1 establishes a 12-month payable term; §9.3 reduces remaining-fee liability to 6 months on termination. The two windows cannot both be operative. The drafting risk is that on early termination, Vendor would claim §9.3 governs (limiting recovery) while continuing to bill the §4.1 term.

03 · SEVERITY

HIGH · sign-blocking by default. Severity is defended against your playbook rubric, not our opinion. This finding is sign-blocking because it changes commercial obligations on remaining-fee liability — directly affecting renewal accounting and early-termination exposure. A “MED” classification would require a playbook entry that explicitly downgrades intra-section fee inconsistencies; you do not have one.

04 · ACTION

Escalate to Legal. Redline §9.3 to mirror §4.1's term. Drafted in `REDLINES.docx`: replace “six (6)” with “twelve (12)” in §9.3, with comment citing §4.1 mirror. Expected counter from Vendor; defensible position is that §9.3 must mirror §4.1's term.

05 · OWNER

Suggested: General Counsel. Owner field is editable in the XLSX. CFO loop-in recommended given the ~€100K delta on a €200K annual contract under early-termination scenario.

08

ARCHITECTURE & CLOSING

Five stages. No hidden seventh.

The pipeline from PDF to register. Same architecture for sample contracts and live engagements; same audit log; same data-handling envelope.

01 · INGEST	02 · DETECT	03 · VALIDATE	04 · CALIBRATE	05 · SHIP
OCR & parse	Rules + LLM	5-element check	Playbook overlay	Register + memo
PDF → clause-graph. Cross-references resolved. OCR-failed pages flagged as Cat 7.	DET / LLM-C / HYB detectors fire in parallel. Candidate findings with confidence scores.	Citation · explanation · severity · action · owner. Misses drop to low-confidence queue.	Severity mapped to client rubric. Below-rubric findings dropped. Audit-logged.	XLSX · MEMO.pdf · REDLINES.docx · 45-min founder-led readout.
PDF in	→	→	→	Register out

Data handling envelope

- **Residency.** Azure EU · Frankfurt region · ephemeral processing tenants.
- **Encryption.** TLS 1.3 in transit · AES-256 at rest · per-engagement keys.
- **Retention.** 30-day default · 7-day on request · delete-on-demand · contractual.
- **Training.** We do not train on your data. Pre-signed DPA covers processor obligations.
- **Audit log.** Tenant-scoped · every action recorded · exportable on request.

Mateusz Szczepański
PRINCIPAL ENGINEER · OPSSOLVED

OPS-METH-BRIEF · V2.0
OCT 2025 · 13 / 13 · END