Better Call CLAUSE: A Discrepancy Benchmark for Auditing LLMs Legal Reasoning Capabilities

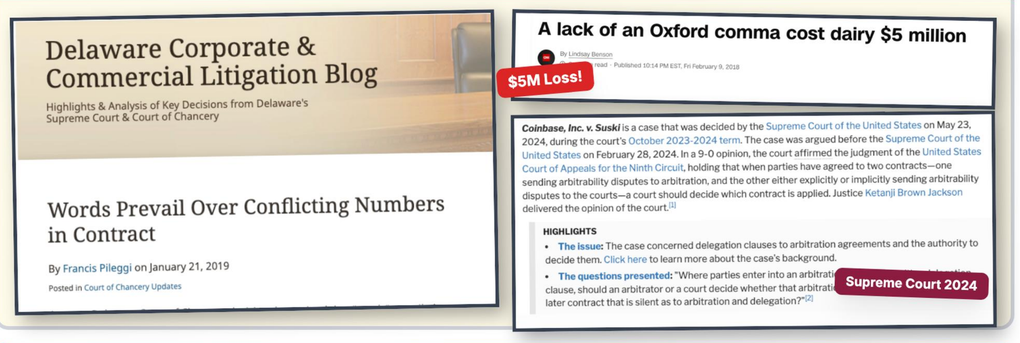
Manan Roy Choudhury[1]*, Adithya Chandramouli[1]*, Mannan Anand[1]*, Vivek Gupta[1]

[1]Arizona State University • *Equal Contribution

EACL 2026 RABAT • MOROCCO
Mars • March 24-29, 2026 • مارس
CoRAL

Scan for Paper

Hidden Errors in Contracts Cost Millions



- ⚠ **High Stakes:** Single omitted clause → $14.5M liability
- **Blind Spots:** LLMs miss subtle legal contradictions
- **The Gap:** No benchmark for real contract flaws

10 Perturbation Types Across 2 Dimensions

Outer-Law (Statutory Violations)

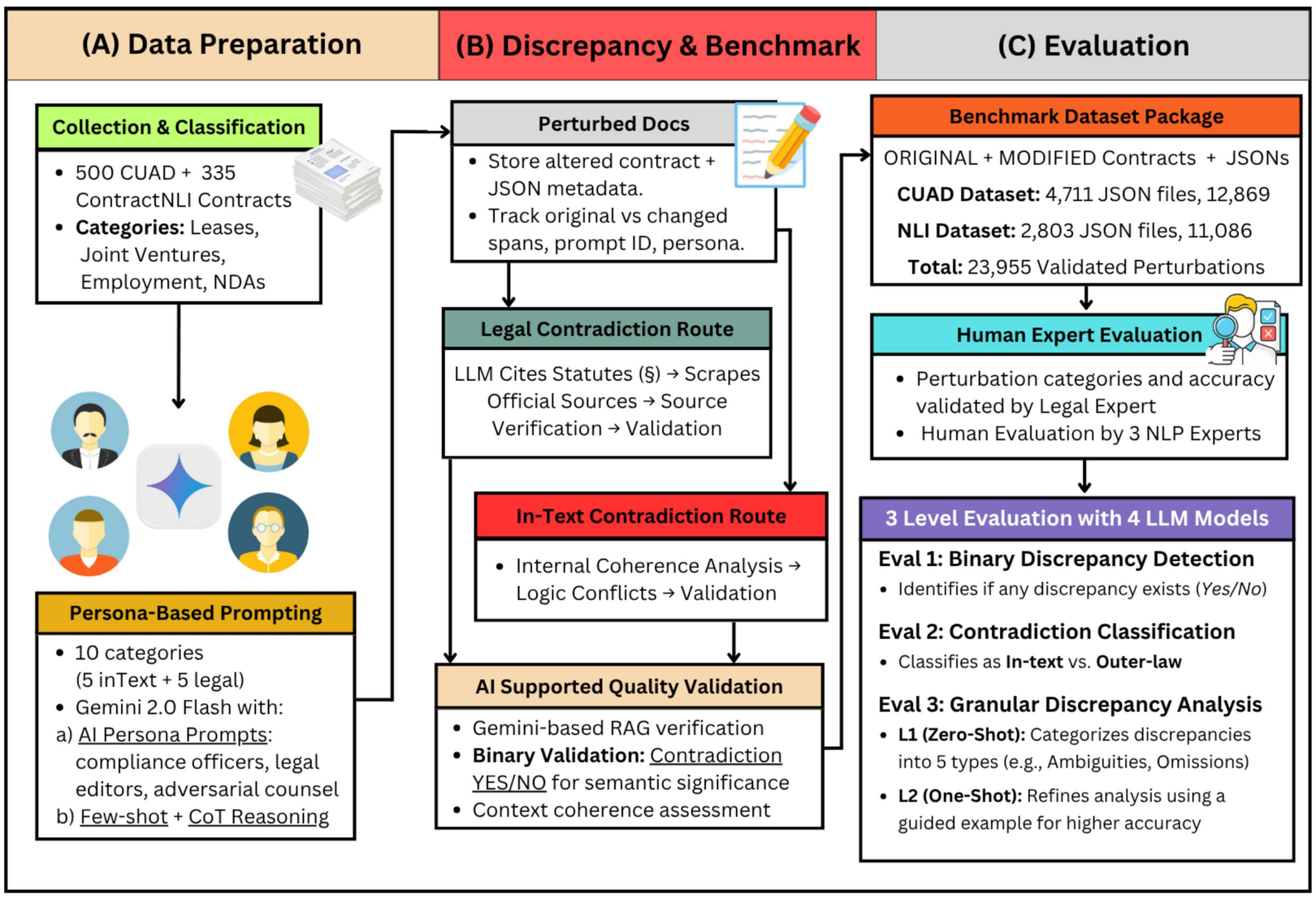
Ambiguity	Vague compliance
Omission	Missing waivers
Misaligned Term.	Statutory deviation
Structural Flaws	Buried disclosures
Inconsistencies	Illegal deadlines

In-Text (Internal Contradictions)

Ambiguity	Conflicting defs
Omission	Missing refs
Misaligned Term.	Undefined terms
Structural Flaws	Broken cross-refs
Inconsistencies	Mutual exclusion

23,955 Total Perturbations | **7,514** Contract Files | **98.58%** Expert Validation | **0.98κ** Inter-Annotator

The CLAUSE Generation & Validation Pipeline



Dataset Breakdown

23,955 Total Perturbations

CUAD Dataset	12,869 (54%)
ContractNLI	11,086 (46%)

4.9K	5.1K	5.5K	4.6K	3.9K
Ambig	Incon	MisT	Omis	StrFl

3-Level Hierarchical Evaluation

- **Eval 1: Binary Detection** — Is there a discrepancy? (Yes/No)
- **Eval 2: Classification** — In-Text vs Outer-Law category
- **Eval 3: Granular Analysis** — 5-way type + explanation quality

Models: GPT-4o, Gemini 2.0, Gemini 2.5, LLaMA 3.3

Human Expert Validation

98.58% Validation Rate | **0.98** Cohen's κ

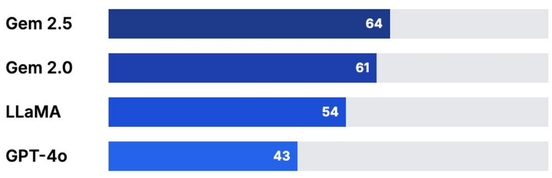
3 NLP Experts • Legal Domain Review

Prompting Strategies Tested

Zero-Shot (Baseline)	Few-Shot (3 Examples)
CoT (Chain-of-Thought)	Persona (Legal Expert)

Key Findings: LLMs Struggle with Legal Reasoning

Fig 1. Detection F1 Score (%)



Best model misses 36% of issues

Fig 2. Miss Rate vs Extra Rate (%)

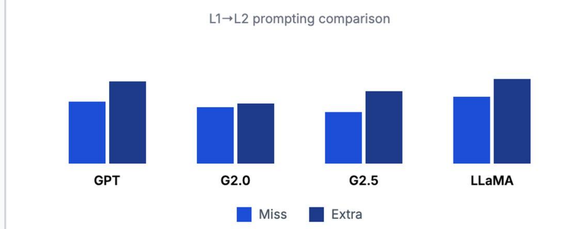
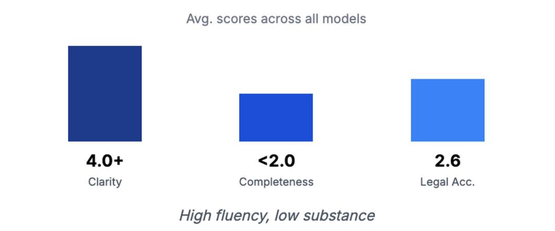


Fig 3. Explanation Quality (0-5)



High fluency, low substance

Fig 4. Law Match Accuracy

<14%

Maximum Achieved

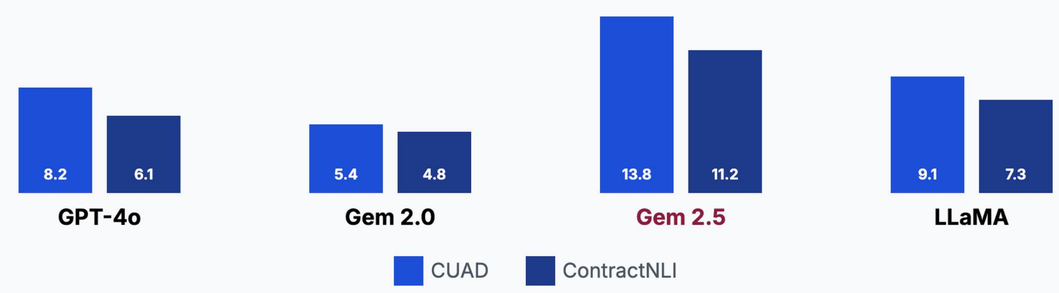
Models fail to cite correct statutes

Binary Detection F1 Scores (CUAD Dataset)

Category	GPT-4o	Gemini 2.0	Gemini 2.5	LLaMA 3.3
Ambiguity	41.2	60.8	**63.7**	53.5
Inconsistency	42.5	60.5	**63.8**	53.9
Misaligned Term.	45.4	**63.5**	63.2	53.9
Omission	41.8	59.7	**63.8**	52.2
Structural Flaw	37.9	50.9	**64.2**	6.9

Best scores in bold. LLaMA fails on Structural Flaws (6.9%, underlined).

Law Match Accuracy by Model & Dataset (%)



Even the best model (Gem 2.5) matches <14% of law citations

IMPLICATIONS

1) Benchmark Creation

CLAUSE delivers the first comprehensive, human-validated dataset containing **23,955** nuanced legal perturbations specifically designed to stress-test adversarial legal reasoning.

2) AI Readiness

Despite impressive linguistic fluency, today's LLMs completely lack the depth of external legal reasoning and citation grounding required for unassisted deployment in high-stakes contract review.