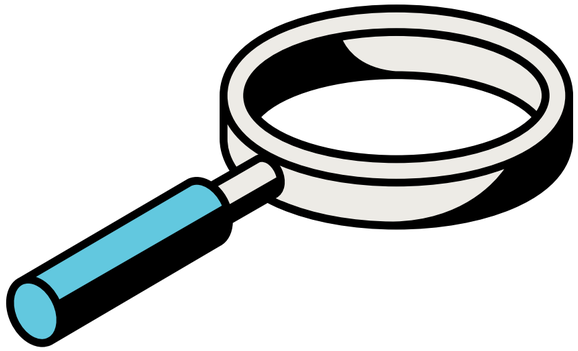
EACL 2026
RABAT • MOROCCO
Mars • March 24-29, 2026 • مارس

CORAL

Scan for Paper

Better Call CLAUSE

A Discrepancy Benchmark for Auditing LLMs Legal Reasoning Capabilities



Presented by: Mannan Anand*
Manan Roy Choudhury* , Adithya Chandramouli*, Dr. Vivek Gupta

*Equal Contribution

WHY THIS MATTERS?

AI in Law

Promise and Precision Problems

- AI tools are rapidly entering legal fields like contract review.
- But legal documents are incredibly precise – tiny changes can have huge consequences.

The Questions:

- Can AI actually detect when something legally important is made ambiguous (or contradictory)?
- Can AI reliably determine if a contract term potentially conflicts with relevant laws or statutes?
- Can it explain why it matters, like a legal expert would?
- And, how well do current AI systems actually perform at these complex legal reasoning tasks when rigorously tested?

MOTIVATION

Case Studies

Words Prevail Over Conflicting Numbers in Contract

By Francis Pileggi on January 21, 2019

Posted in Court of Chancery Updates

A recent Delaware Court of Chancery decision determined that "words" prevailed over "numbers" when they appear next to each other as contract terms in a manner that is inconsistent and contradictory. In *Fetch Interactive Television, LLC v. Touchstream Technologies, Inc.*, C.A. No. 2017-0637-SG (Del. Ch. Jan. 2, 2019), the court described in extensive detail the claims and counterclaims between parties who had entered into licensing agreements and a sublicense for a patent, amid deteriorating personal and business relationships.

A lack of an Oxford comma cost dairy $5 million

By Lindsay Benson

2 minute read · Published 10:14 PM EST, Fri February 9, 2018

Coinbase, Inc. v. Suski is a case that was decided by the Supreme Court of the United States on May 23, 2024, during the court's October 2023-2024 term. The case was argued before the Supreme Court of the United States on February 28, 2024. In a 9-0 opinion, the court affirmed the judgment of the United States Court of Appeals for the Ninth Circuit, holding that when parties have agreed to two contracts—one sending arbitrability disputes to arbitration, and the other either explicitly or implicitly sending arbitrability disputes to the courts—a court should decide which contract is applied. Justice Ketanji Brown Jackson delivered the opinion of the court.[1]

HIGHLIGHTS

- **The issue:** The case concerned delegation clauses to arbitration agreements and the authority to decide them. Click here to learn more about the case's background.
- **The questions presented:** "Where parties enter into an arbitration agreement with a delegation clause, should an arbitrator or a court decide whether that arbitration agreement is narrowed by a later contract that is silent as to arbitration and delegation?"[2]

Our 2-Step Approach

STEP 1

Creating the "Test Track" (Benchmark Construction):

- First, we built a unique testing ground. Using over 500 real contracts, we systematically introduced 10 types of realistic, tricky legal issues that lawyers often watch for (like ambiguities, omissions, or inconsistencies) in 2 flavors (legal and in-text).

- A key step was performing an internal quality check on these embedded issues, especially potential law conflicts (using automated lookups), to ensure our benchmark presents valid and meaningful challenges.

Our 2-Step Approach

STEP 2

“Grading” the AI (Performance Evaluation):

- With this benchmark ready, we then test how well different AI models handle these challenges.

- We measure their performance clearly: Did they find the right issue? Was their legal explanation sound? Did they succeed overall?

  This helps answer the question: How capable is legal AI today?

Contracts Used

Contract Understanding Atticus Dataset (CUAD)

Dataset | *Publication* | *Code* | *Contributors*

A dataset of legal contracts with rich expert annotations

- *13,000+ labels*
- *510 contracts*
- *41 categories of clauses*

Contract Understanding Atticus Dataset (CUAD) v1 is a corpus of 13,000+ labels in 510 commercial legal contracts that have been manually labeled under the supervision of experienced lawyers to identify 41 types of legal clauses that are considered important in contact review in connection with a corporate transaction, including mergers & acquisitions, etc.

CUAD is curated and maintained by The Atticus Project, Inc. to support NLP research and development in legal contract review.

Read the full CUAD v1 announcement here!

Dataset

Version 1	README/Datasheet	License
CUAD v1	Download here.	CC BY 4.0

Publication

Our paper on CUAD is accepted by NeurIPS 2021, the 35th Conference on Neural Information Processing Systems (Datasets and Benchmarks Track)!

OUR CATEGORIZATION

Contractual Pitfall	Brief Description	Key U.S. Case Example	What Exactly Went Contractually Wrong	Core Consequence
Ambiguity	Terms are unclear, vague, or open to multiple reasonable interpretations.	*Fetch Interactive Television, LLC v. Touchstream Technologies, Inc.*	A cure period was stated as "fifteen (30)" days, creating a direct textual contradiction and ambiguity in the timeline.	Disputed deadlines, unexpected obligations based on court interpretation.
Omission	Failure to include necessary clauses, definitions, schedules, or legally mandated terms, or the accidental deletion of protective language.	*Perini Corp. v. Greate Bay Hotel & Casino, Inc.*	The construction contract lacked a waiver of consequential damages or a clear limitation of liability for such damages.	Exposure to unforeseen and potentially catastrophic claims for consequential or punitive damages (in Perini, multi-million dollar lost profits award).
Inconsistencies	Different parts of the contract (or related agreements) stipulate conflicting duties, timelines, definitions, or dispute resolution mechanisms for the same performance or subject matter.	*Coinbase, Inc. v. Suski*	Two separate agreements between the same parties (User Agreement and Sweepstakes Rules) contained conflicting dispute resolution clauses—one delegating arbitrability to an arbitrator, the other a forum selection clause for courts.	Uncertainty, breach of contract claims, difficulty in determining the true obligation, and necessity for court intervention to resolve which term or agreement governs.

CONTINUED...

Contractual Pitfall	Brief Description	Key U.S. Case Example	What Exactly Went Contractually Wrong	Core Consequence
Misaligned Terminology	The same term is used to mean different things in various parts of the contract, different terms are used for the same concept inconsistently, or the scope of a defined party (e.g., "Company," "Affiliate") is unclear or varies.	*Lithko Contracting, LLC, et al. v. XL Insurance America, Inc. (by analogy)*	A waiver of subrogation clause used "no party shall be liable to another party" but then inconsistently referred to "Tenant [i.e., Amazon]" separately, creating ambiguity as to whether Amazon was included as "a party" for all purposes of the waiver within the contractual documents.	Indemnity, waivers, or other critical clauses may not cover the intended parties or apply as expected, leading to disputes over scope of protection.
Structural Flaws	The organization or placement of clauses within a contract, or the relationship between different contractual documents, creates contradictions, obscures legally required disclosures, or renders terms difficult to enforce.	*Carnival Cruise Lines, Inc. v. Shute (by analogy for clause placement and fairness)*	A forum selection clause was included in a form contract (cruise ticket), and its placement and communication were scrutinized for fundamental fairness to determine enforceability.	Legally significant clauses may be deemed unenforceable if not reasonably communicated or if their placement is fundamentally unfair, especially in consumer contracts; uncertainty regarding which terms prevail.

1: Making The Benchmark

1 INPUT

We utilize real-world legal contracts (from the CUAD dataset).

2 METHOD

We employ AI (Gemini) guided by specific 'persona' prompts (e.g. "counsel") to introduce 10 distinct types of realistic perturbations.

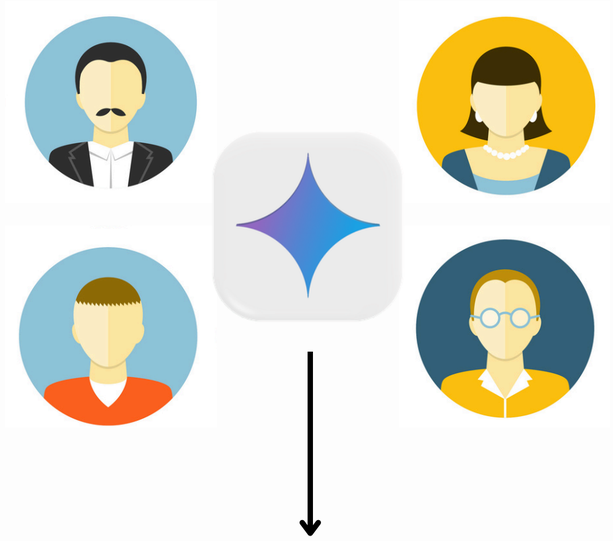
3 VALIDATION (INTERNAL)

- **In-Text Issues:** Where clauses conflict within the same document. Checked for internal inconsistency.
- **Legal Issues:** Where a clause potentially conflicts with external statutory law. Checked against scraped law text for potential conflict.

4 OUTPUT

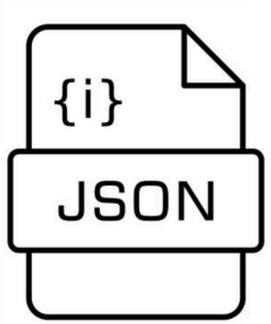
The final benchmark comprises the perturbed contracts paired with detailed (JSON) metadata (tracking changes, explanations, validation scores, etc.).

What We Produce



AI-Generates Perturbations

Perturbed Contracts



Metadata

Prioritized Online Sources for Law Text:

- **Federal Government:** Sites ending in `.gov` or `.mil`.
- **Specific Federal Resources:**
  - US Code via `uscode.house.gov`.
  - Code of Federal Regulations via `ecfr.gov`.
  - Govt. Publishing Office via `govinfo.gov`.
- **State Governments:** Official state sites (e.g., `*.state.az.us`, `*.state.ca.us`).
- **Authoritative Legal Institute:** Cornell Law School's Legal Information Institute (`law.cornell.edu`).

What We Produce

The Dataset is AI Generated

We want to make sure that the validation is correct, which is where we need your expertise.


```
"file_name": "MetLife,Inc.-RemarketingAgreement.txt",
"perturbation": [
  {
    "type": "Ambiguities - Ambiguous Legal Obligation",
    "original_text": "(b) The Remarketing Agents agree (i) to use commercially reasonable efforts to remarket the Remarketed Securities tendered or deemed tendered to the Remarketing Agents in the Remar
    "changed_text": "(b) The Remarketing Agents will endeavor (i) to exert efforts to remarket the Remarketed Securities tendered or deemed tendered to the Remarketing Agents in the Remarketing, (ii) to
    "explanation": "The original text contains a clear obligation for the Remarketing Agents to use \"commercially reasonable efforts\" and \"to notify promptly\". The modified text uses ambiguous terms
    "contradicted_law": "Breach of Contract - Good Faith and Fair Dealing",
    "law_citation": "NY UCC § 1-304",
    "law_url1": [
      "https://www.law.cornell.edu/ucc/1/1-304"
    ],
    "law_url2": [
      "https://codes.findlaw.com/ny/uniform-commercial-code/ucc-sect-1-304.html"
    ],
    "law_explanation": "The implied covenant of good faith and fair dealing in New York requires parties to a contract to act honestly and reasonably in their dealings. The modified text replaces clear
    "location": "Section 1(b)",
    "scraped_snippet_1": "§ 1-304. Obligation of Good Faith. | Uniform Commercial Code | US Law | LII / Legal Information Institute Please help us improve our site! × No thank you § 1-304. Obligation of
    "scraped_snippet_2": "New York Consolidated Laws, Uniform Commercial Code - UCC § 1-304 | FindLaw New York Consolidated Laws, Uniform Commercial Code - UCC § 1-304. Obligation of Good Faith Current
    "scrape_success": 2
  },      You, 2 days ago • GOAT Prompts, 2 URL Web Scraping and RAG Checks…
  {
    "type": "Ambiguities - Ambiguous Legal Obligation",
    "original_text": "(f) By no later than 4:30 P.M., New York City time, on the Remarketing Settlement Date, provided that there has been a Successful Remarketing, the Remarketing Agents shall advise,
    "changed_text": "(f) By approximately 4:30 P.M., New York City time, on or around the Remarketing Settlement Date, assuming there has been a seemingly Successful Remarketing, the Remarketing Agents
    "explanation": "The original text sets a clear deadline and mandatory action to advise specific parties. The modified text introduces uncertainty with \"approximately,\" \"on or around,\" \"assuming,
    "contradicted_law": "Breach of Contract - Failure to Provide Notice",
    "law_citation": "NY CLS UCC § 2-607",
    "law_url1": [
      "https://codes.findlaw.com/ny/uniform-commercial-code/ucc-sect-2-607.html"
    ],
    "law_url2": [
      "https://www.nysenate.gov/legislation/laws/UCC/2-607"
    ],
    "law_explanation": "Under NY CLS UCC § 2-607, if a buyer (here, metaphorically the 'Company' as a recipient of services) accepts tender of goods (or services), the buyer must notify the seller (the
    "location": "Section 1(f)",
    "scraped_snippet_1": "New York Consolidated Laws, Uniform Commercial Code - UCC § 2-607 | FindLaw New York Consolidated Laws, Uniform Commercial Code - UCC § 2-607. Effect of Acceptance; Notice of
    "scraped_snippet_2": "NYS Open Legislation | NYSenate.gov Sorry, you need to enable JavaScript to visit this website. Skip to main content Legislation Search OpenLegislation Statutes Search Term
    "scrape_success": 2
  },
```

Benchmark Quality

Before including any potential issue (a "perturbation") in our final benchmark dataset, we perform an automated internal check to increase confidence that it represents a plausible problem worth testing an AI on.

- **For Internal Inconsistencies:** We use an automated check to see if the change we introduced appears to logically conflict with another specific part of the *same* contract document.
- **For Potential Legal Conflicts:** Similarly, we perform an automated check to assess if the change seems to conflict with relevant external law (often involving looking up cited statutes online and comparing texts).
- **Outcome:** Both types of checks result in a simple internal **"YES / NO" plausibility flag** for each potential issue. This flag helps us filter our dataset during creation, prioritizing meaningful and challenging scenarios for the final benchmark.

Yes? There is a contradiction.
No? Then there isn't an explicit contradiction

Examples

Original Contract Snippet (Section 18.B):

"...Subcontractor expressly assumes the risk of loss or injury that may result from the Work.

B. **Subcontractor shall comply with all applicable laws** (including, without limitation the Federal Occupational Safety and Health Act, Hazardous Communication Requirements, and all applicable environmental protection laws, rules, regulations and ordinances), ordinances, rules, regulations and lawful orders of any public authority having jurisdiction for the safety of persons or property or to protect them from damage, injury or loss. Subcontractor shall comply with all occupational safety and health requirements, including such related publications (not included; but incorporated herein by reference):
· National Electrical Code Handbook, most recent edition. ..."

Modified Contract Snippet (Section 18.B):

"...Subcontractor expressly assumes the risk of loss or injury that may result from the Work.

B. <*p*>**Subcontractor shall make reasonable efforts to comply with applicable laws**<*p*> (including, without limitation the Federal Occupational Safety and Health Act, Hazardous Communication Requirements, and all applicable environmental protection laws, rules, regulations and ordinances), ordinances, rules, regulations and lawful orders of any public authority having jurisdiction for the safety of persons or property or to protect them from damage, injury or loss. Subcontractor shall comply with all occupational safety and health requirements, including such related publications (not included; but incorporated herein by reference):
· National Electrical Code Handbook, most recent edition. ..."

Legal Contradiction


```
file_name: "FTENETWORKS,INC_02_18_2016-EX-99.4-STRATEGICALLIANCEAGREEMENT.txt"
▾ perturbation: [] 3 items
  ▾ 0:
      type: "Ambiguities - Ambiguous Legal Obligation"
      original_text: "Subcontractor shall comply with all applicable laws (including, without limi
        Act, Hazardous Communication Requirements, and all applicable environmental protection laws
        rules, regulations and lawful orders of any public authority having jurisdiction for the sa
        om damage, injury or loss."
      changed_text: "Subcontractor shall endeavor to comply with what they perceive to be applicab
        Occupational Safety and Health Act, Hazardous Communication Requirements, and all applicabl
        ns and ordinances), ordinances, rules, regulations and lawful orders of any public authorit
        r property or to protect them from damage, injury or loss."
      explanation: "Changing 'shall comply' to 'shall endeavor to comply with what they perceive t
        cretion, weakening the obligation to adhere to all laws."
      contradicted_law: "Occupational Safety and Health Act (OSHA)"
      law_citation: "29 CFR § 1926.20(b)(1)"
    ▾ law_url1: [] 1 item
        0: "https://www.osha.gov/laws-regs/regulations/standardnumber/1926/1926.20"
    ▾ law_url2: [] 1 item
        0: "https://www.govinfo.gov/app/details/CFR-2011-title29-vol8/CFR-2011-title29-vol8-sec1926
      law_explanation: "OSHA requires strict compliance with safety standards. The modified text m
        ory nature of OSHA regulations and potentially leading to unsafe working conditions. The o
        w ALL laws. Modifying the language to only require the subcontractor follow what they *per
        o properly enforce."
      location: "18. B."
  ▸ 1:
  ▸ 2:
```

Examples

Original Contract Snippet (Section 4.1.3):

"...Section 4.1.2 shall be (i) the adjustment to the promotion fee as set forth in Section 6.1.2 and (ii) the termination right set out in Section 12.2.2.

4.1.3 Target Incentive Compensation. In addition, [] *and continuing throughout the remainder of the Term, Valeant shall ensure the incentive compensation package for each Sales Representatives requires that at least fifty percent (50%) of the target incentive compensation is derived from achieving target sales of the Product. On at least a quarterly basis, the Parties will meet, through the JSC, to review the target incentive compensation and the actual incentive compensation paid out to the Sales Representatives to discuss, in good faith, any appropriate adjustments to the sales targets and goals related to the Product (but not to the above-mentioned fifty percent (50%) threshold of the target incentive compensation), with the intent of achieving, on average, an actual payout to the Sales Representatives of 50% of their incentive compensation relating to sales of the Product.**

4.1.4 Alliance Managers. Each Party shall appoint a person who shall oversee interactions between the Parties for all matters related to this Agreement..."

Modified Contract Snippet (Section 4.1.3):

"...Section 4.1.2 shall be (i) the adjustment to the promotion fee as set forth in Section 6.1.2 and (ii) the termination right set out in Section 12.2.2.

<*p*>4.1.3 Target Incentive Compensation. Valeant shall determine the incentive compensation package for each Sales Representatives. The Parties will meet, through the JSC, to discuss, in good faith, any appropriate adjustments to the sales targets and goals related to the Product. <*p*>

4.1.4 Alliance Managers. Each Party shall appoint a person who shall oversee interactions between the Parties for all matters related to this Agreement..."


```
file_name: "DovaPharmaceuticalsInc_20181108_10-Q_EX-10.2_11414857_EX-10.2_
▼ perturbation: [] 3 items
  ▶ 0:
  ▼ 1:
      type: "Omissions - In Text Contradiction"
      original_text: "4.1.3 Target Incentive Compensation. In addition, [***]
        ure the incentive compensation package for each Sales Representatives
        pensation is derived from achieving target sales of the Product. On a
        eview the target incentive compensation and the actual incentive comp
        h, any appropriate adjustments to the sales targets and goals relatec
        eshold of the target incentive compensation), with the intent of achi
        of their incentive compensation relating to sales of the Product."
      changed_text: "4.1.3 Target Incentive Compensation. Valeant shall deter
        The Parties will meet, through the JSC, to discuss, in good faith, ar
        Product."
      explanation: "By omitting the clause mandating that at least 50% of the
        n contradicts the intended financial incentive structure. This omissi
        int Steering Committee (JSC) is tasked to decide on the acceptable fc
        nsation matters, which assumed the previously defined compensation st
      location: "ARTICLE 4"
      contradicted_location: "Section 3.3.7"
      contradicted_text: "3.3.7 decide on the acceptable form of and review a
        ers described in Section 4.1.3, including any applicable adjustments
        ives;"
  ▶ 2:
```

Experimental Evaluation & Results

A rigorous, multi-tiered framework to audit the fragile legal reasoning capabilities of state-of-the-art LLMs.

The Three-Tiered Evaluation Framework

Eval 1: Binary Detection

A zero-shot classification task asking models to determine if a legal document contains *any* form of discrepancy (Yes/No).

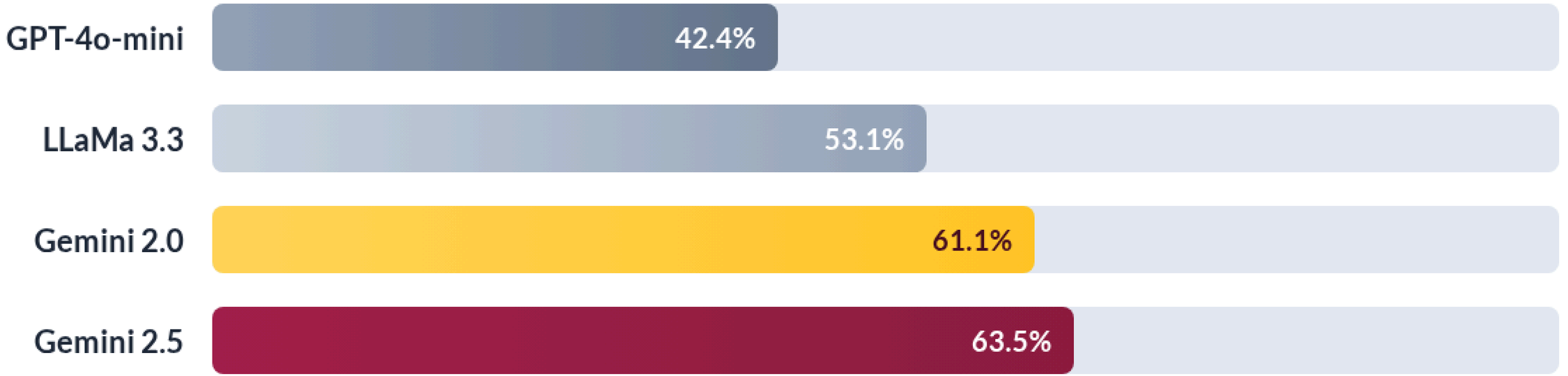
Eval 2: Classification

Categorizing the detected discrepancy strictly into one of two types: an internal **In-Text** contradiction or an external **Outer-Law** contradiction.

Eval 3: Granular Analysis

The ultimate test: models must identify the exact discrepancy span, generate a sound natural language explanation, and cite the violated law.

Eval 1: Binary Discrepancy Detection (F1-Scores)



Key Finding: *Gemini 2.5 leads driven by a high-recall strategy (90%+), but overall precision across all models remains remarkably low. Identifying legally significant absences (Omissions) proved exceptionally challenging.*

Eval 2: Contradiction Type Classification Accuracy

Discrepancy Category	GPT-4o-mini	Gemini 2.0	Gemini 2.5	LLaMa 3.3
Ambiguity	60.8%	61.5%	50.3%	52.7%
Inconsistencies	60.1%	64.2%	51.2%	53.6%
Structural Flaws	60.2%	55.6%	47.6%	51.3%
Misaligned Terminology	55.7%	59.7%	49.7%	53.1%
Omissions	59.4%	57.6%	52.4%	50.8%

Eval 3: The Quality of Legal Explanations

4.1 / 5

Average Clarity Score

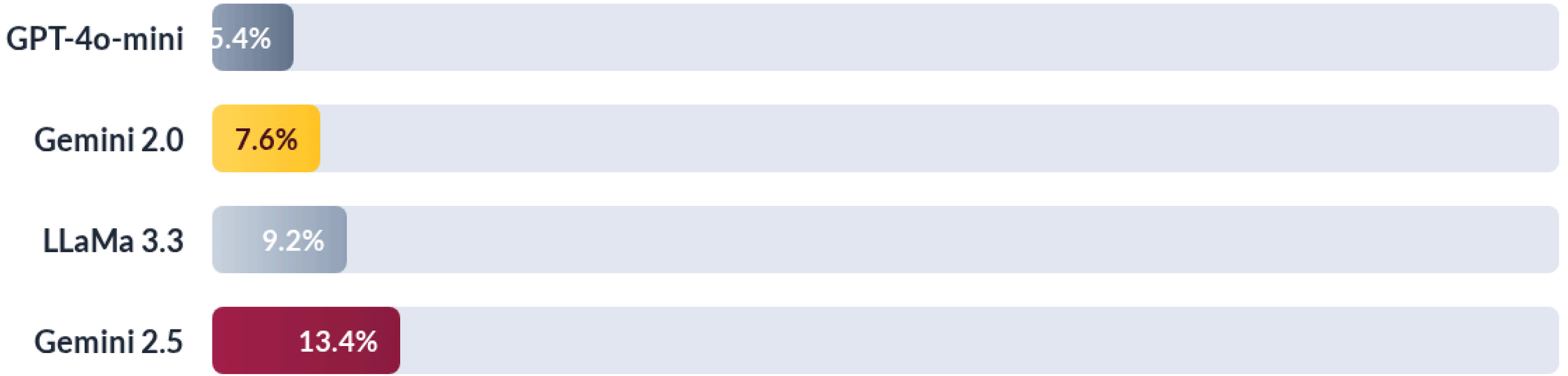
Models generate highly fluent, easy-to-read, and structurally sound explanations. They project immense confidence.

1.9 / 5

Average Completeness

A severe dissociation: models articulate *what* changed, but completely fail to grasp *why* it matters legally or the specific risks it creates.

Eval 3: Semantic Law Match Capabilities



Critical Failure Mode: *Extracting and semantically matching precise legal citations is extremely limited. Even the best-performing model (Gemini 2.5) fails to correctly match external statutory law citations in over 86% of cases.*

Key Takeaways & Conclusion

- **Struggle with Precision:** While models can achieve high recall, they generate significant false positives, especially when utilizing one-shot (L2) prompting.
- **Fluent but Shallow Reasoning:** LLMs produce incredibly articulate legal explanations that mask a profound lack of substantive legal depth and completeness.
- **Failure in Legal Grounding:** The ability to accurately retrieve and match relevant external laws and statutes is currently a major bottleneck for all tested architectures.
- **Not Ready for Unsupervised Deployment:** Despite rapid advancements, today's out-of-the-box LLMs lack the rigorous legal reasoning required for high-stakes, real-world contractual practice.