A Two-Dimensional Evaluation Framework for Factual and Reasoning Assessment of LLMs in Legal Question Answering

Sinan Gultekin[1] Matteo Rossi Reich[1] Francesca Galloni[1]
Francesca Lagioia[1,4] Elena Consiglio[2] Giovanni Sartor[1,4]
Sara Bagnato[3]

Affiliations:

- [1] CIRSFID Alma-AI, Faculty of Law, University of Bologna, Italy
- [2] University of Palermo, Italy
- [3] University of Roma LUMSA, Italy
- [4] European University Institute, Law Department, Italy

Corresponding Authors: Elena Consiglio (elena.consiglio@unipa.it) and Francesca Lagioia (francesca.lagioia@unibo.it)

Why Current LLM Evaluation Falls Short in Legal Domains

Surface-Level Metrics

Standard metrics (BLEU, ROUGE, BERTScore) capture surface-level similarity, not legal correctness

Expert Limitations

Expert human evaluation is resource-intensive, idiosyncratic, and prone to cognitive load decline

Critical Gap

LLMs can produce **legally correct answers through hallucinated or unsound reasoning**

Traditional evaluation metrics miss this critical distinction

Why It Matters: Deploying LLMs for high-stakes legal tasks (asylum proceedings, contract review, legal research) requires trustworthy evaluation beyond binary correct/incorrect.

Real-World Domain: Evidence-Based Legal Analysis

Use Case - Italian Asylum Proceedings

Dataset:

- 91 asylum cases from Tribunal of Palermo (2014-2023)
- Appeal decisions from Territorial Commissions in Western Sicily
- Four outcome types: Refugee (19), Humanitarian (31), Subsidiary (21), Rejected (20)

Document Types:

- C3 Form (standardized intake questionnaire)
- Interview Transcripts (personal hearing records)
- Territorial Commission Initial Decision
- Tribunal Judgment (first-tier appeals)

Why Asylum Law?

Requires evidence-based reasoning and fact extraction

High stakes: human lives and legal status

Representative of complex real-world legal analysis

Good testbed for evaluating reasoning quality

Evaluation Pipeline Architecture

Methodology Overview

01

Questions & Guidelines Design

Bottom-up approach

- Extract **"interesting"** information **regardless of legal relevance**
- Draft evaluation guidelines
- Iteratively refine based on expert feedback

02

Ground-Truth Construction

- 4 legal experts independently
- Inter-evaluation agreement: 75.7%

03

Information Extraction

- LLMs extract answers + provide reasoning grounded in source text

04

Answer Evaluation

Expert + LLM-judges evaluate using 2D framework

05

Performance Comparison

Measure LLM-judge accuracy against expert assessments

Our Methodology

Dimension 1 - Evaluating Answer Accuracy

For N/A Answers:

- **True Negative:** Information not in document (correct N/A)
- **Missing Answer:** Information exists but model said N/A (error)

For Definite Answers (YES/NO or open):

- **Correct:** Accurately matches ground truth (ignoring minor phrasing)
- **Partial:** Incomplete or mixes correct + irrelevant/incorrect info
- **Wrong:** Contradicts ground truth or entirely unrelated

Our Methodology

Dimension 2 - Evaluating Reasoning Quality

Incoherent

the model failed to understand the question

Hallucination

Invented information absent from source document

Inference

Conclusions not necessarily entailed by text (but consistent with it)

Sound

Based on text, clear understanding, no fabrication

Example: The 2D Framework in Action

Question:

What relationship did the applicant have with their family members? What is it now?

LLM Answer:

"In the past, they lived with their father, mother, and three sisters. Currently, they have lost contact with the mother, and there is no information about the sisters."

LLM Reasoning:

The text states: 'To have lived with my family, consisting of my father and mother. I have three younger sisters.' and 'To have no more news of my mother since the day she was taken by the rebels'.

Ground Truth:

They lived with the father and mother; current relationship is not mentioned.

2D Evaluation:

- **Truthfulness:** PARTIAL (added information about sisters not in source)
- **Reasoning:** HALLUCINATION (invented sister information)

Single Metric Would Miss:

Model appears to answer, but introduces fabricated details

Example: The 2D Framework in Action

Question:

Has the applicant suffered abuse / violence / threats / physical and/or psychological violence / intimidation / or been forced into prostitution? (YES, NO, N/A)

LLM Answer:

Yes

LLM Reasoning:

The text states: “my mother told me that the police came looking for me several times.” This implies direct action by the authorities toward the applicant, which can be interpreted as a form of intimidation and a potential threat.

Ground Truth:

Yes

2D Evaluation:

- **Truthfulness:** CORRECT (The answer matches the ground truth)
- **Reasoning:** HALLUCINATION (invented information about the police and directed action by the authorities toward the applicant)

Single Metric Would Miss:

Model appears to answer, but introduces fabricated details

Example: The 2D Framework in Action

Question:

Under what circumstances and for what reasons did the applicant decide to come to Italy?

LLM Answer:

"He decided to come to Italy in 2007 out of fear of suffering ill-treatment due to his sexual orientation, after learning that his partner had been assaulted."

LLM Reasoning:

"The text states: 'he declared that he left Mali in 2007 solely out of fear of suffering ill-treatment because of his sexual orientation; ... after he had learned that his partner had been assaulted."

Ground Truth:

He decided to come to Italy in 2007 out of fear of suffering ill-treatment due to his sexual orientation, after learning that his partner had been assaulted.

2D Evaluation:

- **Truthfulness:** CORRECT
- **Reasoning:** CORRECT

Stage 1 - Extracting Answers from Legal Documents

LLM Information Extraction

Prompting Strategy: Few-Shot Chain-of-Thought

- 5 carefully selected examples
- Mix of YES/NO and open-ended questions
- Examples directed toward difficult questions

Two Prompt Variants Tested:

Prompt 1

Three debating lawyers reaching majority consensus

Result: 89.81% Truthfulness Accuracy

Prompt 2

Disputatio-style debate (one argues for answer, one argues for N/A)

Result: 78.69% Reasoning Soundness Accuracy

Selected for deployment due to better reasoning evaluation

Adversarial framing improves reasoning evaluation despite slight truthfulness trade-off.

Comparing Prompt Strategies

Information Extraction Performance

Performance with P.1	Accuracy	F1 Score
Mistal 7B	65.3%	0.56
Qwen 2.5 7B	59.3%	0.55
Overall	62.3%	0.55
Gemma 3 27B	72.1%	0.71
o4-mini	**82.4%**	**0.80**
Overall	77.3%	0.75

Performance with P.2

Mistal 7B	66.2%	0.57
Qwen 2.5 7B	65.7%	0.56
Overall	65.9%	0.57
Gemma 3 27B	69.5%	0.66
o4-mini	**81.8%**	**0.79**
Overall	75.7%	0.73

Comparing Prompt Strategies

Downstream Judgment Performance

QA Model	EM Ratio	Overall F1	Truth. Acc.	Truth. F1	Reason. Acc.	Reason. F1	Avg. Time
G. 27b P.1	63.33%	0.92	88.78%	0.89	71.09%	0.71	22'43"
o4-mini P.1	74.17%	0.94	90.83%	0.91	80.58%	0.81	22'43"
P.1 Overall	68.75%	0.94	**89.81%**	**0.90**	75.83%	0.76	22'43"
G. 27b P.2	65.64%	0.93	87.95%	0.88	73.85%	0.74	15'17"
o4-mini P.2	74.36%	0.95	87.88%	0.89	83.53%	0.84	28'46"
P.2 Overall	**70.00%**	**0.94**	87.92%	0.88	**78.69%**	**0.79**	**22'02"**

Stage 2 - Automated Evaluation

LLM-as-a-Judge Framework

Two-Step Evaluation Process:

Truthfulness Assessment

Reasoning Soundness Assessment

Addressing Evaluator Biases in LLM-as-a-Judge

Bias Mitigation Strategies

1

Positional & Verbosity Bias

- Pointwise, reference-based evaluation (not pairwise comparison)
- Eliminates structural preference for specific answer positions/lengths

2

Overconfidence Bias

- Disputatio mechanism (adversarial debate)
- Model forced to consider counterarguments
- Improves truthfulness through debate dynamics

3

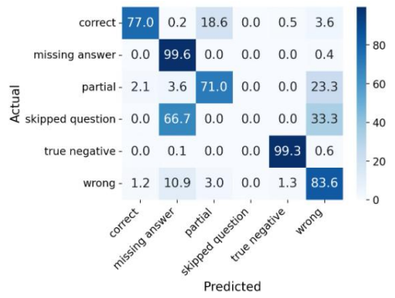
Self-Enhancement Bias

- Judge model ≠ generator model
- Architectural separation prevents vested interest
- Promotes objective assessment

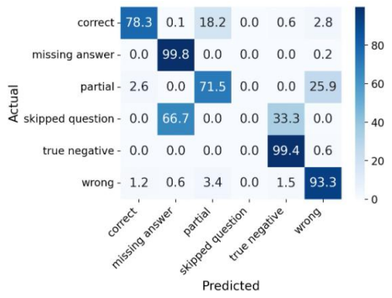
LLM-as-a-Judge Performance Across Models

Judge Model	EM Ratio	Truth. Accuracy	Truth. F1 Score	Reasoning Accuracy	Reasoning F1 Score	Avg. Time per Case
Gem. 2.5 F.	71.4%	**90.5%**	**0.92**	77.8%	0.80	**26'45"**
Gem. 2.5 F. L.	70.3%	89.6%	0.92	77.0%	0.80	142'48"
Gpt-5 Nano	**74.5%**	89.8%	0.92	**82.6%**	**0.81**	132'31"
Mistral S. 24B	53.8%	88.4%	0.90	56.0%	0.65	32'20"
Qwen 3 14B	69.9%	87.5%	0.90	77.5%	0.79	58'15"

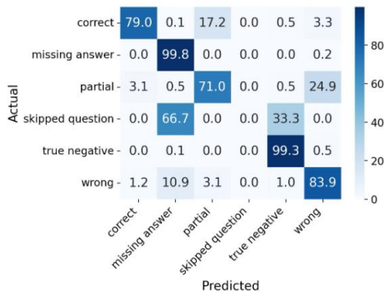
Confusion Matrices Results



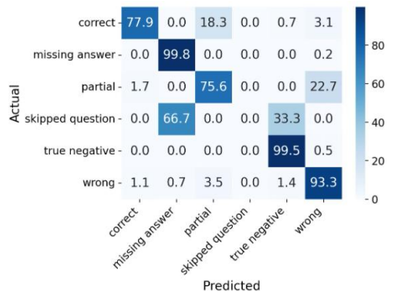
Qwen 14B



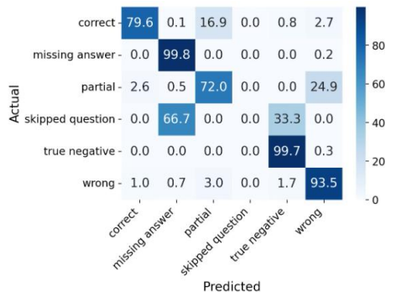
GPT-5 Nano



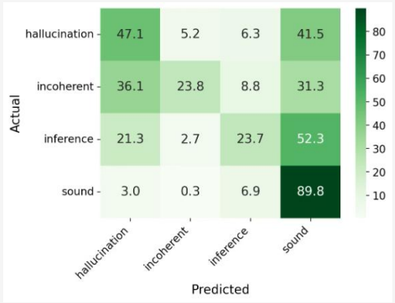
Mistral 24B



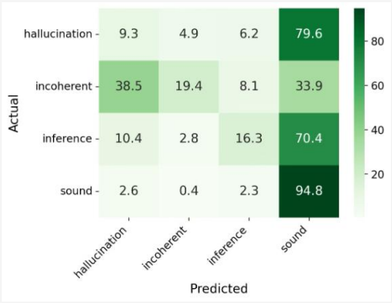
Gemini Flash Lite



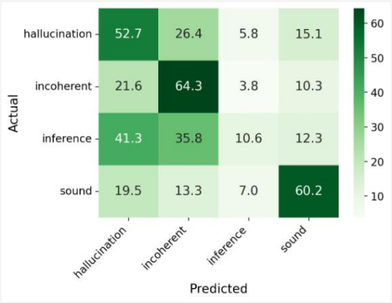
Gemini Flash



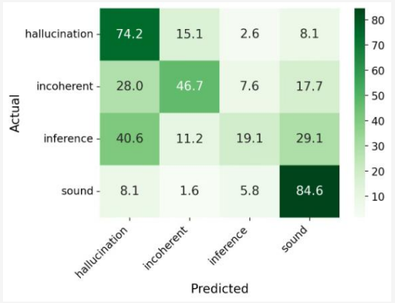
Qwen 14B



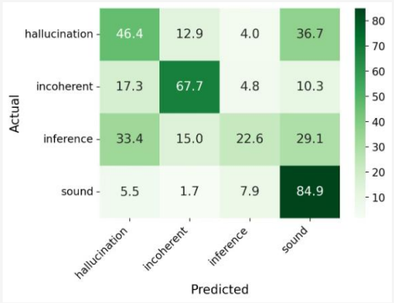
GPT-5 Nano



Mistral 24B



Gemini Flash Lite



Gemini Flash

Future Works

Apply framework to other complex legal tasks

Combine multiple models to use their respective strengths

Thank You

This work was partially supported by the following projects: CompuLaw – Computable Law – funded by the ERC under the Horizon 2020 (Grant Agreement N. 833647); PRIN2022 PRIMA - PRivacy Infringements Machine-Advice (Ref. Prot. n.: 20224TPEYC - CUP J53D23005130001); PRIN2022 EQUAL – EQUitableALgorithms (Ref. Prot n. 2022KFLF3E_001 - CUP J53D23005560001); "FAIR - Future Artificial Intelligence Research" – Spoke 8 and Spoke 1 (CAI4DSA action) under the European Commission's NextGeneration EU programme, PNRR – M4C2 – Inv. 1.3, Partenariato Esteso (PE00000013).