Legal Knowledge and Information Systems
R. Markovich et al. (Eds.)
© 2025 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA251579

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

Sinan GULTEKIN [a], Matteo ROSSI REICH [a], Francesca GALLONI [a], Francesca LAGIOIA [a,d,1], Elena CONSIGLIO [b,1], Giovanni SARTOR [a,d], Sara BAGNATO [c]

[a] *CIRSFID Alma-AI, Faculty of Law, University of Bologna, Italy*
[b] *University of Palermo, Italy*
[c] *University of Roma LUMSA, Italy*
[d] *European University Institute, Law Department, Italy*

Abstract. Deploying Large Language Models (LLMs) for legal question-answering requires ensuring factual accuracy and logical coherence. Current evaluation metrics inadequately capture legal reasoning complexity, while expert assessments lack scalability. We propose a two-dimensional framework that independently measures Truthfulness and Reasoning Soundness in model outputs, applied to Italian asylum proceedings requiring evidence-based analysis. This dual-axis approach reveals critical issues—such as legally correct answers derived through unsound or hallucinatory reasoning—that standard metrics fail to detect. To enable large-scale application, we implement an automated LLM-as-a-Judge system with bias-mitigation techniques. Experimental results demonstrate strong correspondence between automated judgments and expert evaluations, confirming framework reliability. This work advances diagnostic methodology for assessing LLMs in legal domains, offering both theoretical insight and practical applicability toward more trustworthy and accountable legal AI systems.

Keywords. LLM, prompting, LLM-as-a-judge, Legal AI

1. Introduction

Large Language Models (LLMs) attract growing interest for automating legal question answering (QA), assisting practitioners in extracting information from large volumes of unstructured documents to support scalable decision-making. Yet, the legal field imposes stringent demands on accuracy, explainability, and transparency, which LLMs do not consistently guarantee [1], making their evaluation challenging. Standard automated metrics fail to assess legal correctness or capture reasoning nuances [2], while human

[1] Corresponding Authors: Elena Consiglio at elena.consiglio@unipa.it and Francesca Lagioia at francesca.lagioia@unibo.it

evaluation, though context-sensitive, is resource-intensive, idiosyncratic [3], and prone to declining accuracy under cognitive load [4]. We therefore propose a granular evaluation framework based on the LLM-as-a-Judge paradigm.

Our main contribution is a two-dimensional framework that dissects answers' quality along two dimensions: output accuracy and reasoning soundness. We test such framework on a demanding task, i.e., extracting information from Italian asylum proceedings. This task is representative of real-world practice and well-suited to evaluating the reasoning abilities of LLMs.

This work builds on recent critiques of the LLM-as-a-Judge paradigm. Studies show that even top-performing models exhibit significant shortcomings, especially in areas requiring specialized knowledge [5]. We tackle these limitations by implementing structured prompts, rigorous human-grounded assessments, and a two-dimensional framework applied to information extraction from a dataset of Italian asylum cases. By focusing on legal context, we provide a scalable and reliable evaluation methodology applicable in high-stakes domains, where expert knowledge is essential and extractive QA alone is insufficient to answer a question.

The paper is structured as follows: Section 2 reviews the related works, Section 3 describes the dataset and Section 4 presents the two-dimensional evaluation framework and its implementation. Section 5 reports the experimental results. Finally, Section 6concludes and outlines directions for future research.

2. Related Works

The evaluation of natural language generation (NLG) has significantly evolved, moving from simple lexical comparisons to sophisticated, reasoning-based, assessments. This reflects both the increasing complexity of NLG tasks and the growing capabilities of models. Early automated metrics like BLEU[6], METEOR[7], and ROUGE [8] measured surface-level similarity through n-gram analysis. Though scalable and consistent, they struggle with semantic equivalence and penalize valid paraphrasing, producing weak correlation with human judgments [9].

Subsequent methods incorporated semantic understanding through contextual embeddings. BERTScore [10] enabled synonym recognition via token-level similarity, while BLEURT [11] and COMET [12] improved correlations with human assessments through fine-tuning on rating data. However, these methods may not capture higher-order qualities like logical coherence, factual grounding, or argument soundness.

The emergence of powerful LLMs enables a new paradigm, i.e., models acting as evaluators, a concept commonly known as "LLM-as-a-Judge" [13]. Leveraging advanced instruction-following and reasoning capabilities, LLMs emulate context-sensitive expert analysis. In recent studies, LLMs achieve high agreement with human judgments across tasks, ranging from ranking chatbot responses to evaluating output safety and helpfulness, all at greater scalability and lower cost [13,14].

Recent research emphasizes granular diagnostic evaluations. FActScore [15] pioneered atomic evaluation by verifying single factual claims against knowledge sources, while ARES [16] assessed RAG output faithfulness and contextual relevance. G-Eval [17] employed Chain-of-Thought prompting for step-by-step logic analysis, enhancing transparency, alongside methods targeting logical fallacies [18].

The evalution of LLMs' outputs is critical in the legal domain, which demands impeccable factual and legal accuracy, as well as transparent and verifiable reasoning [19]. Legal NLP made substantial progress in creating domain-specific benchmarks like LegalBench [20], which enables systematically measures various facets of legal reasoning. Yet evaluations still reveal limits. Even state-of-the-art LLMs show instability in answering legal questions [21] and the analyses often remain brief and unreliable, even where they follow basic structures like IRAC (Issue, Rule, Analysis, Conclusion), [22].

Most critically, hallucinations remain widespread [23], with rates between 59% and 88% for some queries, casting doubts on the readiness of LLMs for legal practice. While fine-tuning methods are proposed to improve accuracy [24], they usually address either grounding or reasoning in isolation.

Our contribution introduces a two-dimensional evaluation framework that concurrently assesses both truthfulness and reasoning soundness. We operationalize this framework through a specialized LLM-as-a-Judge system, tailored to the evidence-based analysis required in the legal field, providing a granular and diagnostic tool for assessing LLMs' reliability in high-stakes areas, where errors may carry serious consequences.

3. Dataset

As mentioned, in this study, we focus on applications for international protection in Italy. The dataset consists of legal and administrative texts from 91 asylum cases appealed to the Tribunal of Palermo and decided between 2014 and 2023. Of these, asylum was granted in 19 cases, humanitarian protection in 31, subsidiary protection in 21, and 20 cases were fully rejected. All personally identifiable information has been removed.

The procedure begins with an asylum claim, followed by identification and completion of the C3 Form questionnaire. The Territorial Commission evaluates claims and may: 1) recognize refugee status or subsidiary protection; 2) forward paperwork for humanitarian/special protection; or 3) reject the claim. Rejected or unsatisfactory decisions may be appealed to the judicial authority.

The dataset[2] contains decisions appealed to the Tribunal of Palermo, that is competent to adjudicate decisions of Territorial Commissions in Western Sicily (Palermo, Trapani, Agrigento). Representativeness is limited: the dataset contains only decisions from Western Sicily Commissions that were rejected or unsatisfactory, as authorities denied broader access and only cooperating law firms provided cases under appeal. Additionally, most cases involve male applicants, reflecting broader refugee flow patterns but limiting female asylum seeker analysis. The documentation for each case file is composed of one or more of the following documents:

- C3 Form (Modulo C3): A standardised questionnaire detailing the applicant's personal and family information, normally compiled by the police;
- Interview Transcript (Audizione): A verbatim record of the personal interview with the staff of the Territorial Commission, during which the applicant states the grounds for their claim, often conducted with the help of an interpreter. The interview might be recorded and the transcript might be obtained, using AI, from a video recording;

[2] Available on GitHub at https://github.com/EquitableAlgorithms/Two-Dimensional-Evaluation-Framework

- Initial Decision by the Territorial Commission (Commissione): The official ruling, stating the determination on the asylum claim and detailing the reasons for the decision.
- First-Tier Tribunal's Judgment (Tribunale): The judgment from the court of first instance (Tribunale), confirming or revising the initial determination by the administrative authority, with reasons.

Starting from PDF scans, we process these files using Surya [25], a transformer-based OCR toolkit, to create a machine-readable corpus. Given the highly sensitive and personal nature of the contained information, we apply a rigorous anonymisation protocol, to remove all Personally Identifiable Information (PII).

4. Methodology

Our methodology is designed to create a robust, scalable and fine-grained evaluation of LLMs performance in the legal area. As shown in Figure 1, the workflow includes the following steps.

The first step is *Questions and Guidelines Design*.[3] After examining a large set of cases, for each type of document, we identified relevant questions. To avoid biases, we adopted a bottom-up approach: questions were meant to extract any "interesting" information, regardless of its legal relevance to the decision of the case. Additionally, we drafted a set of guidelines to be used by legal experts in answering the same questions. Both guidelines and questions were iteratively refined based on experts' feedback.

The second step is *Ground-truth Construction*. For each case, four legal experts independently answered all questions.

The third step is the *Information Extraction*. For each case, LLMs were asked to address the mentioned questions, providing two outputs: (a) the answer and (b) the reasoning justifying such answer, grounded in the text of the decision.

The fourth step is the *Answer Evaluation*. Both, experts and LLM-judges, evaluated the LLMs answers.

The last step is *Performance comparison.* We assessed the LLM-judges evaluations against those of experts, to determine the former's performance.

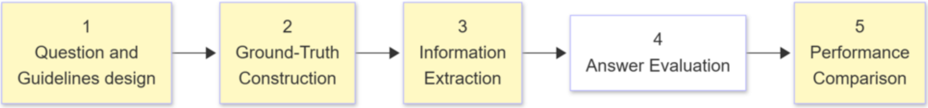


Figure 1. Evaluation Pipeline

4.1. The Methodology for Evaluating the LLMs Output

The evaluation protocol aims to ensure a consistent and nuanced evaluation of LLMs-generated answers[4]. It proceeds along two primary dimensions: *Truthfulness* and *Reasoning Soundness*, as shown in Figure 2.

[3] The question set is available on GitHub at https://github.com/EquitableAlgorithms/Two-Dimensional-Evaluation-Framework

[4] The evaluation guidelines are available on GitHub at https://github.com/EquitableAlgorithms/Two-Dimensional-Evaluation-Framework

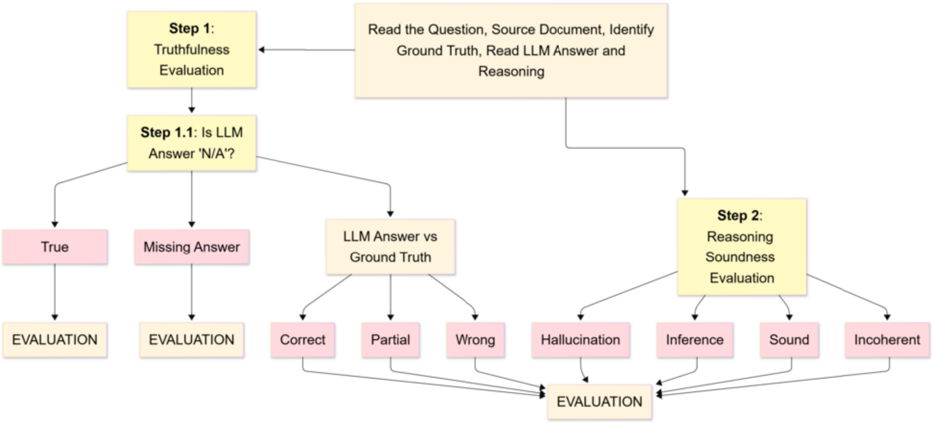

Step 1: Truthfulness Evaluation
Read the Question, Source Document, Identify Ground Truth, Read LLM Answer and Reasoning
Step 1.1: Is LLM Answer 'N/A'?
True
Missing Answer
LLM Answer vs Ground Truth
Step 2: Reasoning Soundness Evaluation
EVALUATION
EVALUATION
Correct
Partial
Wrong
Hallucination
Inference
Sound
Incoherent
EVALUATION

Figure 2. Evaluation Workflow

The *Truthfulness Evaluation* assesses the answer by LLMs, distinguishing between ‘N/A’ (information is Not Available in the text) answers, and those providing a definite response to a YES/NO or an open question. ‘N/A’ answers are labelled as follows: ‘*True Negative*’ if the requested information is absent from the document; or ‘*Missing Answer*’ if it is present. Definite answers are compared against the Ground Truth, and labelled as follows: ‘*Correct*’ if it accurately matches the Ground Truth, regardless of minor phrasal differences; ‘*Partial*’ if it is incomplete or mixes correct with irrelevant or incorrect information; ‘*Wrong*’ if it contradicts the Ground Truth or provides entirely unrelated information.

The *Reasoning Soundness Evaluation* labels the LLMs’ *reasoning* as follows: ‘*Incoherent*’ if it is nonsensical or completely off-topic; ‘*Hallucination*’ if it incorporates invented information, absent in the source document; ‘*Inference*’ if it consists in inferences, leading to conclusion that are not necessarily entailed by the text (while being consistent with such a text and even somehow supported by it): ’*Sound*’ if it is based on the text, with a clear understanding of the question and without fabrication.

The combination of Truthfulness and Reasoning Soundness provides an analytical assessment of the quality and reliability of information extraction, as shown in Table 1.

A crucial aspect of our methodology pertains to ensuring the correctness and consistency of the ground truth. Thus, legal experts proceed according to the following steps: (i) thoroughly understand the question, (ii) review the source, (iii) establish the Ground Truth, and (iv) examine the LLMs’ answers and their reasoning.

After evaluating an initial set of cases, experts were given an alignment batch (939 question–answering and reasoning items) to compute the inter-evaluation agreement and solve doubts. Given the open-ended formulation of questions, the complexity of documents and the diversity of contents, with many inherently ambiguous borderline cases, the agreement rate of 75.7% was satisfactory. Thus, we proceeded with the distribution of the remaining cases. Disagreement often arose because several points could be pertinent in a single case, leading to context-dependent interpretations and multiple justifiable reasoning paths. This complexity reflects the richness of legal evaluation rather than inconsistency among annotators, and it highlights the challenge of defining a single, unambiguous ground truth in such domains. Thus, we proceeded with the distribution of the remaining cases

Table 1. Legal Experts Evaluation

Questions	LLM Answer	LLM Reasoning	Evaluation	Rationale	Ground Truth
Q3: What relationship did the applicant have with their family members? What relationship do they have now with their family members?	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.	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'.	Partial, Hallucination	Information about the sisters is not present in the text.	They lived with the father and mother; the current relationship is not mentioned.
Q3.1: Does the commission consider the elements related to problems with family members to be relevant? (YES, NO, N/A)	NO	The text does not indicate that the commission considers the elements related to problems with family members to be relevant. The commission focuses on other aspects of the applicant's case.	Wrong, Inference	The commission does not dwell on these aspects; the question in question cannot be answered.	N/A

4.2. LLMs-Based Information Extraction

For the information extraction, we employ a "Few-shot-CoT" approach [26,27], including 5 examples, 2 YES/NO questions, and 3 open questions. Examples were selected so as to direct the LLM to answer the most difficult questions correctly. We run experiments with smaller and bigger models; the latter enabled a considerable jump in F1score on the information extraction task, as can be seen from Table 2. Therefore, we experiment with 2 prompts on the big models [5] to understand the performance on the downstream task of automatic answer evaluation. The first prompt instructs the LLM to impersonate three debating lawyers who must reach a majority consensus. The second one implements a *disputatio*-style debate, wherein one persona argues for a proposed answer while a second contends that the answer is "N/A". Both prompts, optimized using an LLM [28], ask for a structured JSON output and require the model to provide its *reasoning* for each answer.

As reported in Table 3, the first prompt had a marginally higher Truthfulness Accuracy (89.81% vs. 87.92%). The second prompt reached a significantly higher Reasoning Soundness Accuracy (78.69% vs. 75.83%). Moreover, it demonstrated slightly better overall answer quality, with higher scores in both Exact Match Ratio (70.00% vs. 68.75%) and F1 Score (0.94 vs. 0.93).

We selected the second prompt considering that its superiority in reasoning was crucial for performance evaluation in the LLM-as-a-judge framework.

Table 2. Information Extraction performance

(a) Performance with P.1

LLM QA Model	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

(b) Performance with P.2

LLM QA Model	Accuracy	F1 Score
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

[5] The feature extraction prompts are available on GitHub at https://github.com/EquitableAlgorithms/Two-Dimensional-Evaluation-Framework

Table 3. Prompt techniques for Information Extraction

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.2 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”**

We sent the questions to the models in batches,increase efficiency. An automated validation layer then parsed the outputs to ensure that every question was answered correctly. If a question was missed, a targeted re-prompting mechanism queried the model with the specific unanswered question up to five times, significantly enhancing the system’s robustness and overall reliability.

4.3. LLM-as-a-Judge

To enable evaluation at scale, we developed an “LLM-as-a-Judge” framework [13]. This approach combines scalability with explanatory reasoning, to provide pointwise evaluation of content accuracy, including a final rating and a descriptive feedback.

The prompt for LLM-as-a-Judge includeed our detailed evaluation guidelines to direct the evaluation [6].

A critical challenge in deploying LLMs as evaluators is their susceptibility to biases that can compromise reliability. Our framework incorporated several design choices to mitigate such biases. We addressed presentation-related biases such as **positional** [13,29] and **verbosity** bias [13,30] by employing a pointwise, reference-based evaluation: the LLM-judge assessed each single candidate answer against the corresponding ground truth. This eliminated the structural opportunity for the model to prefer one answer over another based on presentation order or length. We addressed content and cognitive biases through prompt engineering. To counter **authority** bias [30], where seemingly authoritative claims might sway a model, our prompt explicitly instructed the judge to verify all answers against the provided source document. This forced a grounded evaluation, preventing the model from accepting unsubstantiated assertions. Furthermore, the disputation mechanism, which simulates an adversarial debate, reduced **overconfidence** bias by forcing the judge to consider counterarguments, a technique shown to improve truthfulness [31]. Finally, to prevent **self-enhancement** bias, where a model tends to rate its outputs more favorably [32,33], the judge’s model was different from the one producing the output being evaluated. This architectural separation ensures the evaluator has no vested interest in the generated output, promoting a more objective assessment.

According to our guidelines, the evaluation included distinct “explanation-rating” steps to enhance the reliability of the final rating. The pipeline consisted of the following steps. First the LLM checks whether the evaluated model has refused to answer, returning ‘N/A’. For YES/NO or open answers, the system proceeds to a two-step evaluation. First, it assesses the Truthfulness of the answer by comparing it against the ground truth,

[6] The LLM-as-a-judge prompt is available on GitHub at https://github.com/EquitableAlgorithms/Two-Dimensional-Evaluation-Framework

identifying discrepancies, and rendering a verdict. Then it assesses the Reasoning Soundness by matching the LLM's reasoning with both the human-provided reasoning and the whole source documents. This technique facilitates identifying hallucinations [17]. This structured, step-by-step process also enhances the interpretability of the judge's decision-making, ensuring more transparent evaluations.

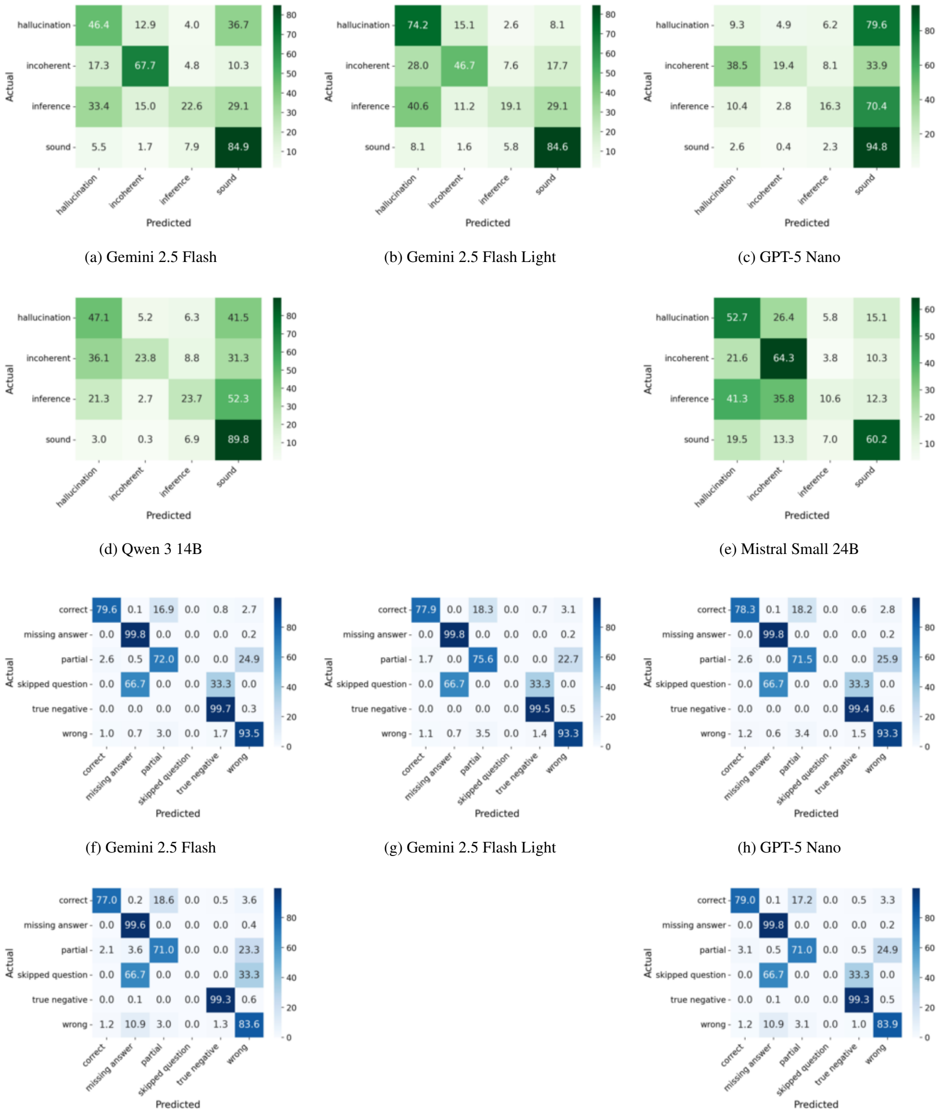
5. Experimental Results

Gemini 2.5 Flash (G2.5 F.) was evaluated on the full dataset as baseline. We compared Gemini 2.5 Flash Lite (G2.5 F. L.), Gpt-5 Nano, Mistral Small 24B 2501, and Qwen3 14B on a stratified subset of 20 cases (each ¿300 questions), enabling the 'Thinking' feature for energy-efficient comparison. The results of this comparative evaluation are presented in Table 4 and Figure 3.

Table 4. A comparison of Judge model performance metrics

Judge Model	EM Ratio	Truth. Accuracy	Truth. F1 Score	Reasoning Accuracy	Reasoning F1 Score	Avg. Time per Case
Gem. 2.5 F. Full	71.3%	90.4%	0.92	77.8%	0.80	**26'45"**
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"

The most immediate observation is the substantial difference in computational cost; enabling the 'thinking' feature increased the average processing time per case up to five-fold, rendering the reasoning approach less practical for scalable deployment. Moreover, all models consistently yielded higher accuracy in Truthfulness than Reasoning Soundness, suggesting that answering is considerably less complex objective than reasoning. While GPT-5 Nano (Thinking) initially appeared to achieve the highest Reasoning Accuracy (82.6%), a deeper analysis of confusion matrices revealed that this is misleading. The model exhibited a strong bias toward the majority class ('sound'), while performing poorly on minority classes (e.g., 'hallucinations'), correctly identifying only 6.2%. In contrast, the Gemini 2.5 Flash models demonstrated a more consistent performance across categories: Gemini 2.5 Flash Light (Thinking) particularly excelled in spotting hallucinations, indicating a more robust, less biased evaluation capability, while the standard Gemini 2.5 Flash provided a superior balance of speed and low computational cost, emerging as the most viable option. On the smaller models side, Mistral-small was outperformed on overall metrics by the rest of its competitors; however, it showed a more balanced performance across the different classes. Even though Qwen3-14B seemed a small, promising alternative to others, a close examination of the confusion matrix and computational time shows that Gemini models predicted classes other than Sound more reliably. A qualitative review of the evaluation process highlighted two significant and recurring challenges. First the LLM Judge generally struggles in differentiating between "inference" and "hallucinations", this happens when the model has to decide if a conclusion is logically derived from premises within the text or if it's new information being incorrectly presented as fact. Second, the consistent application of the 'Partial' label proved difficult due to its inherent subjectivity. The guideline defines this based on "missing crucial information", which forces a value judgment on the part of the human annotator.



(a) Gemini 2.5 Flash

(b) Gemini 2.5 Flash Light

(c) GPT-5 Nano

(d) Qwen 3 14B

(e) Mistral Small 24B

(f) Gemini 2.5 Flash

(g) Gemini 2.5 Flash Light

(h) GPT-5 Nano

(i) Qwen 3 14B

(j) Mistral Small 24B

Figure 3. Confusion matrices for all judge models. **Rows 1-2:** Reasoning Soundness evaluation. **Rows 3-4:** Truthfulness evaluation.

5.1. Computational Cost

Local models were benchmarked on a server equipped with a `Tesla V100-PCIE-32GB` GPU. The inference time per question was influenced by the number of retrials required after a failure to answer, an issue strongly emerging when using the second, more complex prompt. For the first prompt, the `mistralai/Mistral-7B-Instruct-v0.3` and `Qwen/Qwen2.5-7B-Instruct` models recorded inference times of 28.26s and 21.73s, respectively. For the second prompt, these times increased significantly to 90.38s and 150.96s. In contrast, API-based models showed a more contained increase: `google/gemma-3-27b-it` went from 17.84s to 25.76s, and `o4-mini-2025-04-16` remained at 8.26s for both prompts, reflecting their lower retrial rates. The LLM judge, using `gemini-2.5-flash-preview-05-20`, averaged 5.13 seconds per question. Budget is a critical constraint, as each case averages 7.9M input and 3.8M output tokens, making the per-token price of flagship models like Gemini 2.5 Pro prohibitively expensive.

6. Conclusion

We introduced a novel, two-dimensional framework for the granular evaluation of LLMs in legal QA. By distinguishing performance in Truthfulness and Reasoning Soundness, we obtained a more nuanced assessment. We operationalized this framework using an LLM-as-a-Judge system, incorporating adversarial prompting and architectural choices to mitigate evaluative biases. Our experiments, grounded in a real-world dataset of Italian asylum cases, demonstrate the viability of our approach.

Our findings indicate that while current models can assess truthfulness with high accuracy, the evaluation of reasoning soundness remains a significant challenge. Results revealed a performance gap across all tested models, highlighting the complexity of scrutinizing logical chains of thought.

There is a clear need to refine the LLM-as-a-Judge's capacity for evaluating reasoning, potentially through more sophisticated prompting techniques or model fine-tuning for logical analysis. Exploring the framework's applicability to other complex legal tasks, such as case-brief generation or statutory interpretation, is a further promising avenue for extending this research and moving towards more reliable and transparent AI applications in the legal domain.

7. Acknowledgments

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.: 20224TP-EYC - 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)

References

[1] Doyle C, Tucker AD. If You Give an LLM a Legal Practice Guide. In: Proceedings of the 2025 Symposium on Computer Science and Law; 2025. p. 194-205.

[2] Schluter N. The limits of automatic summarisation according to ROUGE. In: Lapata M, Blunsom P, Koller A, editors. Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Valencia, Spain: Association for Computational Linguistics; 2017. p. 41-5. Available from: https://aclanthology.org/E17-2007/.

[3] Clark E, August T, Serrano S, Haduong N, Gururangan S, Smith NA. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In: Zong C, Xia F, Li W, Navigli R, editors. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics; 2021. p. 7282-96. Available from: https://aclanthology.org/2021.acl-long.565/.

[4] Kern C, Eckman S, Beck J, Chew R, Ma B, Kreuter F. Annotation Sensitivity: Training Data Collection Methods Affect Model Performance. In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023. p. 14874-86. Available from: https://aclanthology.org/2023.findings-emnlp.992/.

[5] Szymanski A, Ziems N, Eicher-Miller HA, Li TJJ, Jiang M, Metoyer RA. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In: Proceedings of the 30th International Conference on Intelligent User Interfaces; 2025. p. 952-66.

[6] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a Method for Automatic Evaluation of Machine Translation. In: Isabelle P, Charniak E, Lin D, editors. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics; 2002. p. 311-8. Available from: https://aclanthology.org/P02-1040/.

[7] Rei R, Stewart C, Farinha AC, Lavie A. COMET: A Neural Framework for MT Evaluation. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 2685-702. Available from: https://aclanthology.org/2020.emnlp-main.213/.

[8] Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics; 2004. p. 74-81. Available from: https://aclanthology.org/W04-1013/.

[9] Reiter E. A Structured Review of the Validity of BLEU. Computational Linguistics. 2018 Sep;44(3):393-401. Available from: https://aclanthology.org/J18-3002/.

[10] Zhang* T, Kishore* V, Wu* F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. In: International Conference on Learning Representations; 2020. Available from: https://openreview.net/forum?id=SkeHuCVFDr.

[11] Sellam T, Das D, Parikh A. BLEURT: Learning Robust Metrics for Text Generation. In: Jurafsky D, Chai J, Schluter N, Tetreault J, editors. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 7881-92. Available from: https://aclanthology.org/2020.acl-main.704/.

[12] Rei R, Stewart C, Farinha AC, Lavie A. COMET: A Neural Framework for MT Evaluation. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 2685-702. Available from: https://aclanthology.org/2020.emnlp-main.213/.

[13] Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23. Red Hook, NY, USA: Curran Associates Inc.; 2023. .

[14] Chan CM, Chen W, Su Y, Yu J, Xue W, Zhang S, et al. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. ArXiv. 2023;abs/2308.07201:null.

[15] Min S, Krishna K, Lyu X, Lewis M, Yih Wt, Koh P, et al. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 12076-100. Available from: https://aclanthology.org/2023.emnlp-main.741/.

[16] Saad-Falcon J, Khattab O, Potts C, Zaharia M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In: Duh K, Gomez H, Bethard S, editors. Proceedings of

the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico: Association for Computational Linguistics; 2024. p. 338-54. Available from: https://aclanthology.org/2024.naacl-long.20/.

[17] Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In: Bouamor H, Pino J, Bali K, editors. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023. p. 2511-22. Available from: https://aclanthology.org/2023.emnlp-main.153/.

[18] Cheng Q, Sun T, Zhang W, Wang S, Liu X, Zhang M, et al.. Evaluating Hallucinations in Chinese Large Language Models; 2023. Available from: https://arxiv.org/abs/2310.03368.

[19] Zhong M, Zhang A, Wang X, Hou R, Xiong W, Zhu C, et al.. Law of the Weakest Link: Cross Capabilities of Large Language Models; 2024. Available from: https://arxiv.org/abs/2409.19951.

[20] Guha N, Nyarko J, Ho DE, Ré C, Chilton A, Narayana A, et al. LEGALBENCH: a collaboratively built benchmark for measuring legal reasoning in large language models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23. Red Hook, NY, USA: Curran Associates Inc.; 2023. .

[21] Blair-Stanek A, Durme BV. LLMs Provide Unstable Answers to Legal Questions. ArXiv. 2025;abs/2502.05196. Available from: https://api.semanticscholar.org/CorpusID:276249373.

[22] Peoples L. Artificial Intelligence and Legal Analysis: Implications for Legal Education and the Profession; 2025. Available from: https://arxiv.org/abs/2502.03487.

[23] Dahl M, Magesh V, Suzgun M, Ho DE. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. Journal of Legal Analysis. 2024 06;16(1):64-93. Available from: https://doi.org/10.1093/jla/laae003.

[24] Hu Y, Gan L, Xiao W, Kuang K, Wu F. Fine-tuning Large Language Models for Improving Factuality in Legal Question Answering. In: Rambow O, Wanner L, Apidianaki M, Al-Khalifa H, Eugenio BD, Schockaert S, editors. Proceedings of the 31st International Conference on Computational Linguistics. Abu Dhabi, UAE: Association for Computational Linguistics; 2025. p. 4410-27. Available from: https://aclanthology.org/2025.coling-main.298/.

[25] Paruchuri V, Team D. Surya: A lightweight document OCR and analysis toolkit; 2025. GitHub repository. https://github.com/VikParuchuri/surya.

[26] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22. Red Hook, NY, USA: Curran Associates Inc.; 2022. .

[27] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22. Red Hook, NY, USA: Curran Associates Inc.; 2022. .

[28] Gonen H, Iyer S, Blevins T, Smith N, Zettlemoyer L. Demystifying Prompts in Language Models via Perplexity Estimation. In: Bouamor H, Pino J, Bali K, editors. Findings of the Association for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics; 2023. p. 10136-48. Available from: https://aclanthology.org/2023.findings-emnlp.679/.

[29] Shi L, Ma C, Liang W, Diao X, Ma W, Vosoughi S. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge; 2025. Available from: https://arxiv.org/abs/2406.07791.

[30] Ye J, Wang Y, Huang Y, Chen D, Zhang Q, Moniz N, et al.. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge; 2024. Available from: https://arxiv.org/abs/2410.02736.

[31] Khan A, Hughes J, Valentine D, Ruis L, Sachan K, Radhakrishnan A, et al.. Debating with More Persuasive LLMs Leads to More Truthful Answers; 2024. Available from: https://arxiv.org/abs/2402.06782.

[32] Panickssery A, Bowman SR, Feng S. LLM Evaluators Recognize and Favor Their Own Generations; 2024. Available from: https://arxiv.org/abs/2404.13076.

[33] Li R, Patel T, Du X. PRD: Peer Rank and Discussion Improve Large Language Model based Evaluations; 2024. Available from: https://arxiv.org/abs/2307.02762.