

Rabula: A Benchmark for Evaluating LLMs in Brazilian Legal Tasks *

Eduardo Caruso Barbosa Pacheco¹, Fernanda Mattar Suriani² and Ricardo Ribeiro³

¹ Atlas.IA, Santa Catarina, Brazil

² Lawgorithm, São Paulo, Brazil

³ Advocacia-Geral da União (AGU), Brasília, Brazil

Abstract

Rabula is a benchmark for evaluating large language models (LLMs) in the Brazilian legal domain, addressing current limitations of assessments based solely on multiple-choice legal questions. Built upon the 2024 Brazilian Bar Exam (OAB), it includes four tasks: multiple-choice questions, legal document selection, document drafting, and essay-style legal problem solving. Multiple-choice responses are evaluated against the official answer key, while generative tasks are assessed using an LLM-as-judge method, based on official OAB grading criteria. Each generated answer is scored through a set of binary questions weighted by difficulty. Human evaluation using a golden label (from three legal experts) shows high agreement with the best evaluator model (Cohen's Kappa[14] 79.4) in essay-style task, comparable to the agreement among the human reviewers themselves (Fleiss' Kappa[13] 78.9). In document writing task, however, the agreement was smaller (Cohen's Kappa 64.3) compared with humans (Fleiss' Kappa 76.2). Rabula allows for granular assessment of model competence across legal areas and tasks. As a demonstration, we compare OpenAI's gpt4o-mini with Maritaca.AI's Sabiazinho-3, revealing that performance varies both by legal area and task. This benchmark aims to close the evaluation gap in Portuguese-language LLMs and foster the development of more capable legal models in Brazil.

Keywords

Benchmark, LLM, LLM-as-a-Judge

1. Introduction

The Brazilian Judiciary has made significant progress in adopting artificial intelligence for both administrative and judicial tasks. According to the 2023 AI Projects Dashboard in the Judiciary from the National Council of Justice, there were 140 AI projects in 2023, compared to 111 in 2022, with 80% of courts having at least one AI project in 2023. Generative AI holds relevance due to its potential for drafting legal documents, which requires significant effort from human professionals in both court administration and judicial activities. The same dashboard indicates that 39.7% of courts already use or are implementing Generative AI in judicial activities, while 21.9% do so in administrative tasks. The importance of AI has also caught the attention of the Brazilian government, which announced investments of BRL 23 billion by 2028 to develop the sector. Despite these developments, there has been little effort to create or adapt LLMs specifically for Portuguese or specialize them in the legal domain. The recent surge in publicly available LLMs has not translated into the development of foundational models in Portuguese, especially those tailored for the Brazilian legal field. Notable exceptions include Sabiá-3 from Maritaca.AI [1], Juru [2] (based on Sabiá-2 and specialized in the legal domain), and Tucano [3].

While these projects are commendable, the limited development of Portuguese generative AI specialized in the legal domain presents practical challenges for these initiatives and imposes

* *Proceedings of the First Argument Mining and Empirical Legal Research Workshop (AMELR 2025), June 20, 2025, Chicago, United States*

✉ edu@atlasia.tech (E. Pacheco); fernandasuriani@alumni.usp.br (F. Suriani); ricardosribeiro1976@gmail.com (R. Ribeiro)

🆔 0009-0007-9887-9189 (E. Pacheco); 0009-0000-7667-2944 (F. Suriani); 0000-0002-2413-2468 (R. Ribeiro)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

constraints on conclusions regarding model capabilities. Among the mentioned projects, no benchmark exists for the Portuguese legal domain beyond multiple-choice tests.

For instance, Sabiá-3's legal domain performance was evaluated using two multiple-choice exams. The first, introduced in the Sabiá-2 article [4], is based on 160 multiple-choice questions from the 2023 OAB exam for lawyer qualification. The second, called ENAM, is a qualification exam for law graduates aspiring to judicial careers, consisting of 160 multiple-choice questions [1]. For the legal domain, the only task tested was the ability to answer multiple-choice questions. The model achieved 75.9% accuracy on the first phase of the OAB exams and 64.9% on the ENAM.

Tucano [3], on the other hand, used a dataset of 1,820 questions from the first phase of the OAB exam, as detailed in the article "Passing the Brazilian OAB Exam: Data Preparation and Some Experiments" [5]. Tucano performed poorly, with Pearson product-moment correlation scores between 0.21 and 0.34 depending on model size.

Meanwhile, Juru [2], developed in collaboration with Maritaca.AI, was evaluated using 160 multiple-choice questions and 35 multiple-choice questions from the ENADE exam, designed for law students. Juru achieved 82.5% accuracy on ENADE and 62.2% on the OAB, both involving multiple-choice tasks.

The authors of Juru argue that open-ended question evaluations using LLM-as-judge may exhibit low correlation with human evaluators due to domain-specific vocabulary, knowledge, and the inherent difficulty of evaluating open-ended responses, which are costly to assess. They claim that multiple-choice questions serve as better benchmarks in the context of Juru development.

We diverge from this view for several reasons. First, there is not necessarily low correlation between human evaluators and LLM-as-judge, as variations can be mitigated through decision-making processes involving many techniques to improve performance. Second, the official OAB exam provides a scoring key for essay questions, offering a standardized answer that reduces evaluator (human or machine) discretion and simplifies decision-making to a binary compliance check. Thus, the problem characterization was inaccurate: while the questions are open-ended, the responses are evaluated by a less complex task, due to the binary nature of the criteria, mitigating the criticisms raised. Finally, our experiment on response convergence showed a Fleiss Kapa[13] of 0.7893 and 0.7625 depending on the task among three human evaluators. Comparing evaluator models against the golden standard (majority vote among humans) we found a mean Cohen's Kappa[14] of 0.7941 in the best model in one task and 0.7941 in other.

In addition to justifying the adequacy of LLM-as-judge evaluations, it is worth questioning the argument that multiple-choice tests are inherently more suitable. Although multiple-choice tests measure domain knowledge breadth and have their utility, they do not reflect the daily tasks of legal professionals. Tasks like drafting legal documents, selecting the correct procedural document, or providing legal opinions are assessed in the second phase of the OAB exam and reflects the day-to-day tasks of a lawyer. They are good candidates to compose a benchmark. This paper aims to offer a methodology for evaluating Portuguese-language LLMs in the legal domain, both through objective multiple-choice questions from the first phase of the OAB exam (measuring knowledge breadth) and essay questions from the second phase (measuring practical professional skills). As an example of analysis this benchmark allows, we compare OpenAI's gpt4o-mini with Maritaca.AI's Sabiazinho-3 model. While both perform comparably on multiple-choice questions (with a slight edge for gpt4o-mini), Sabiazinho-3 outperforms in most legal subfields on essay tasks.

2. Methodology

Our methodology involves developing an evaluation framework that assesses both the breadth of legal knowledge and practical legal skills. This is achieved through four tasks: multiple-choice questions, selecting the appropriate legal document for a case, drafting the legal document, and responding to legal case questions in free-text format.

The multiple-choice questions correspond to the first phase of the Brazilian Bar Exam (OAB). For these, we prompt the evaluated models to provide objective answers, which are then compared

against the official answer key prepared by legal experts from Fundação Getúlio Vargas (FGV), the institution responsible for administering the OAB exam. The remaining tasks are evaluated using the LLM-as-a-Judge technique [6], which compares the free-text responses of candidate models to the official answer key in the second-phase tasks. This begins with the creation of a golden label, based on the official grading rubric, by having a responder model generate answers that are then reviewed by three experienced human jurists. This review conducted over 21 legal drafting tasks and 84 essay questions from the second phase of the OAB, following the exam’s official criteria.

This process provides not only the expected answer but also a reference for determining whether a given model response meets the required standard. We then select the best evaluator model based on the highest Cohen’s Kappa agreement between its evaluations and the golden label (majority vote among human reviewers).

Once the evaluator model is selected, three independent instances of it, each prompted with the candidate’s response and the grading criteria, assess each criterion individually. The final judgment is determined by majority vote. Each question or task is scored according to the predefined rubric. These scores are aggregated by legal subject and computed as an overall total, providing a comprehensive measure of each model’s performance across tasks.

2.1. OAB Bar Exam

The benchmark is based on the Brazilian Bar Exam (OAB), developed by legal experts from FGV. Like the U.S. bar exam, it determines who is qualified to practice law in Brazil. The exam occurs three times a year and has two phases. The first consists of 80 multiple-choice questions across various areas of law. Those who pass (with at least 50%) proceed to the second phase, where candidates choose one of seven legal specializations and complete two tasks: drafting a legal document based on a practical case and answering four essay-style questions. The second phase requires a minimum score of 60% and is graded using binary rubric: each response element is either correct and awarded points or incorrect and receives none. Selecting the wrong type of legal document results in an automatic zero on the practical task, making it impossible to pass. The essay responses also follow predefined scoring criteria and represent 50% of the final grade. The following figure 1 illustrate examples of an official first-phase OAB exam and the official answer key used for grading responses in the second phase.

2.2. Data

The Dataset consists of 3 tables: multiple-choice, practical, and discursive. These tables correspond to questions that were manually extracted from the OAB exams administered in 2024. The multiple-choice table contains 240 multiple-choice questions from the first phase of the 2024 OAB exams (OAB 40, 41, and 42) and includes the following fields: id, exam, exam_date, number, question, answer, and cancelled_question. The "practical" table contains 21 practical exam cases distributed across the seven specialization areas of the second phase of the OAB exam. These cases were taken from the 2024 second-phase OAB exams (OAB 39, OAB 40, and OAB 41) and include the following fields: id, exam, exam_date, area, question, answer, criteria, and legal_document. The question field contains the exam prompt, the answer field provides general response guidelines, the criteria field includes the set of scoring criteria and their respective points, and the legal_document field specifies the expected legal document. The "discursive" table contains 84 essay-style questions from the seven specialization areas of the second phase across the three OAB exams administered in 2024. Its structure is the same as the "practical" table, but without the "legal_document" field.



12

Ubirajara é membro de uma comunidade indígena situada em terras regularmente demarcadas, ali vivendo conforme as tradições dos seus ancestrais. Em determinado momento, ele resolveu tentar nova vida em uma cidade brasileira. Sem recursos para dar início a esse projeto, decidiu vender a terra em que habitava desde seu nascimento para um grupo de agricultores, que pretende ali se instalar definitivamente.

Sobre a hipótese narrada, segundo a ordem jurídico-constitucional brasileira, assinale a afirmativa correta.

- (A) Ubirajara somente poderá dispor das terras se a alienação, comprovadamente, atender aos imperativos da ordem econômica brasileira.
- (B) Ubirajara, caso figure como proprietário das terras no registro de imóveis da localidade, poderá aliená-las, assegurado o direito de participação da comunidade no valor da venda.
- (C) Ubirajara não pode efetivar a venda almejada, pois as terras em questão não são passíveis de alienação e nem mesmo de disposição.
- (D) Ubirajara somente poderia alienar as terras após a devida autorização por parte da comunidade indígena, que é a proprietária das terras.

QUESTÃO 4

Jorge foi definitivamente condenado à pena de desacato, fixada em um ano de detenção, em regime aberto. Presentes os requisitos, a pena de reclusão foi substituída por uma pena de prestação pecuniária. Foi autorizado o parcelamento do cumprimento da pena em 12 (doze) prestações iguais e sucessivas.

Após o pagamento de cinco parcelas, Jorge faleceu. A filha de Jorge, Janaina, maior e herdeira de bens deixados pelo falecido, procura você, como advogado(a), informando ter obtido novas provas capazes de comprovar a inocência de seu pai, bem como indagando a respeito da sua responsabilidade pessoal pelo pagamento das parcelas da prestação pecuniária que seu pai não quitou em vida.

Assim, responda às questões a seguir.

- A) As parcelas remanescentes da pena de prestação pecuniária poderão ser cobradas de Janaina? Responda, fundamentadamente, indicando o princípio de Direito Penal aplicável. (Valor: 0,65)
- B) Identifique, de forma justificada, se há meios processuais que legitimem Janaina a comprovar a inocência de Jorge. (Valor: 0,60)

Obs.: o(a) examinando(a) deve fundamentar suas respostas. A mera citação do dispositivo legal não confere pontuação.

Gabarito Comentado

A questão exige do examinando conhecimentos sobre extinção da punibilidade e revisão criminal.

A) Não, diante do princípio da intranscendência da pena ou da responsabilidade pessoal ou personalidade ou intransmissibilidade da pena, a morte do condenado extingue a punibilidade, na forma do Art. 107, inciso I, do CP, ou do Art. 5º, inciso XLV, da CRFB/88 ou Artigo 5, item 3, da Convenção Americana sobre Direitos Humanos o Pacto de San José da Costa Rica (aprova pelo Decreto 678/92).

B) Janaina pode pleitear a revisão criminal em favor de seu pai, na forma do Art. 621, inciso III, ou o Art. 623, ambos do CPP.

Distribuição dos Pontos

ITEM	PONTUAÇÃO
A) Não, diante do princípio da intranscendência da pena <u>ou</u> da responsabilidade pessoal <u>ou</u> personalidade <u>ou</u> intransmissibilidade da pena (0,20), a morte do condenado extingue a punibilidade (0,35), na forma do Art. 107, inciso I, do CP, <u>ou</u> do Art. 5º, inciso XLV, da CRFB/88 <u>ou</u> Artigo 5, item 3, da Convenção Americana sobre Direitos Humanos o Pacto de San José da Costa Rica (aprova pelo Decreto 678/92) (0,10).	0,00/0,20/0,30/0,35 0,45/0,55/0,65
B) É cabível a revisão criminal mesmo após a morte da pessoa condenada (0,50), com base no Art. 621, inciso III, <u>ou</u> o Art. 623, ambos do CPP (0,10).	0,00/0,50/0,60

Figure 1: Examples of question of OAB first phase (left) and second phase (right) in Portuguese.

2.3. Data processing

To enable the systematic answering of questions by different models instructed with response prompts, so that they can be evaluated by models instructed with evaluation prompts considering all assessment criteria, we processed the grading criteria for each question in the second phase of the discursive and practical exams using a prompt that converts the criteria, originally presented as continuous text in the official answer key, into JSON format. For example, consider the answer key for Question 1-A of Administrative Law from OAB 41, which appears in the official answer key as follows (translated into English): “A. No. The granting of retirement constitutes a complex act that will only be complete after a ruling by the Court of Accounts (0.50), in accordance with Article 71, item III, of the Brazilian Federal Constitution of 1988 (CRFB/88) or Binding Precedent No. 3 (0.10). 0.00/0.50/0.60” The formatted response in JSON format appears as follows (translated into English):

```
{ "letter": "A", "part": "I", "answer": "No. The granting of retirement constitutes a complex act that will only be complete after a ruling by the Court of Accounts, in accordance with Article 71, item III, of the Brazilian Federal Constitution (CRFB/88) or Binding Precedent No. 3.", "criteria": "No. The granting of retirement constitutes a complex act that will only be complete after a ruling by the Court of Accounts", "points": 0.5 }, { "letter": "A", "part": "II", "answer": "No. The granting of retirement constitutes a complex act that will only be complete after a ruling by the Court of Accounts, in accordance with Article 71, item III, of the Brazilian Federal Constitution (CRFB/88) or Binding Precedent No. 3.", "criteria": "in accordance with Article 71, item III, of the Brazilian Federal Constitution (CRFB/88) or Binding Precedent No. 3", "points": 0.1 }
```

With this format, the models can see both the criterion itself, which is scored (criteria, points), and the answer, which serves as context for the criterion. Considering the sum of individual criteria, the benchmark consists of 379 discursive criteria, 561 criteria in writing practical cases, 21 criteria in legal document identification and 240 multiple choice questions, summing up 1201 evaluation points.

We employed LLMs in various data processing tasks. First, as mentioned, involves formatting the scoring criteria for the discursive and practical questions in the second phase. Next, we designed prompts to generate model responses for evaluation. This category includes four different prompts, one for each task: answering multiple-choice questions, identifying the appropriate legal documents, drafting legal documents, and responding to discursive questions. To organize and system-

atize these prompts, we used the LangChain framework in Python. Finally, we created evaluation prompts, where we applied the LLM-as-a-Judge technique using the "Evaluation by Criteria" approach, which assesses responses based on reference criteria in a binary decision framework (compliance or non-compliance with each criterion).

2.4. LLM-as-a-Judge

It is possible to use LLMs to evaluate responses generated by other models or by humans. The method known as LLM-as-a-Judge is an approach for automated evaluation, which mitigates the high costs of human assessments and enables large-scale evaluations. One study [7] demonstrated that "the result of LLM evaluation is consistent with the results obtained by expert human evaluation: the texts rated higher by human experts are also rated higher by the LLMs." Another study [8] characterized the type of evaluation conducted in this work as "Solving Yes/No Questions" and provided an in-depth discussion on the relevance of using LLMs as judges, highlighting the challenges, ways to overcome them, and the biases present in both LLMs and human evaluators.

A third study [9] found that the level of agreement between humans and LLM-as-a-Judge (using GPT-4) is around 85%, approximately the same level of agreement found among human evaluators (81%). A fourth study [10], from early 2023, demonstrated that ChatGPT (GPT-3.5-turbo) achieved "state-of-the-art or competitive correlation with human judgments in most cases" even in tasks where previous generative models had failed in the field of linguistics. Another study [11], from late 2024, introduced ways to improve LLM-as-a-Judge performance, such as requiring reasoning for the response, a technique we implemented in the present work. Additionally, a 2024 study [12] found that larger models, such as GPT-4 and Llama3-70B, achieve human alignment above 80%, while also highlighting some concerns, such as the higher variance observed in smaller models.

2.5. Alignment with Human Evaluators

To evaluate the alignment between the responses of the evaluator model and human assessments, we conducted the following experiment. First, we instructed an arbitrary model, gpt4o-mini, to respond to two discursive tasks from the benchmark: drafting the legal document in 21 cases and answering 84 discursive questions. Combined, these two tasks provide 940 evaluation points, each consisting of binary questions assessing whether the established criteria in the answer key were met.

Next, we developed an annotation tool using Streamlit and asked 3 experienced legal experts to evaluate the model's responses using this tool. The tool presents the question posed to the model, the answer key, the candidate's response (in this case, the model's response), and the list of relevant criteria for that question. When clicking on a criterion, the user sees a binary question regarding the fulfillment of that criterion, with "yes" or "no" options. The human evaluators answered the questions, and we then asked an LLM-as-a-Judge to respond to the same questions. By comparing the responses, we can assess two key aspects: the agreement between human evaluators (see **Figure 2**) and the agreement between humans and the LLM-as-a-Judge (see **Figure 3**). We calculated Fleiss' Kappa to assess inter-rater agreement among the human evaluators and obtained a score of 0.789. According to the widely adopted interpretation proposed by Landis and Koch[15], this value falls within the range of 0.61 to 0.80, which indicates substantial agreement. This level of consistency among the evaluators provides a reliable foundation for constructing the golden labels used in our benchmark and for comparing model-generated responses. Next, we prompted seven different candidate models (gpt4o, gpt4o-mini, gpt4.1, gpt4.1-mini, Sabiá-3, Sabiazinho-3, gpto3-mini) to act as evaluators using the LLM-as-a-Judge framework, assessing the responses generated by the responder model (gpt4o-mini). For the open-ended essay question task from the second phase of the exam, we ran 25 evaluation executions for each candidate model across all grading criteria.

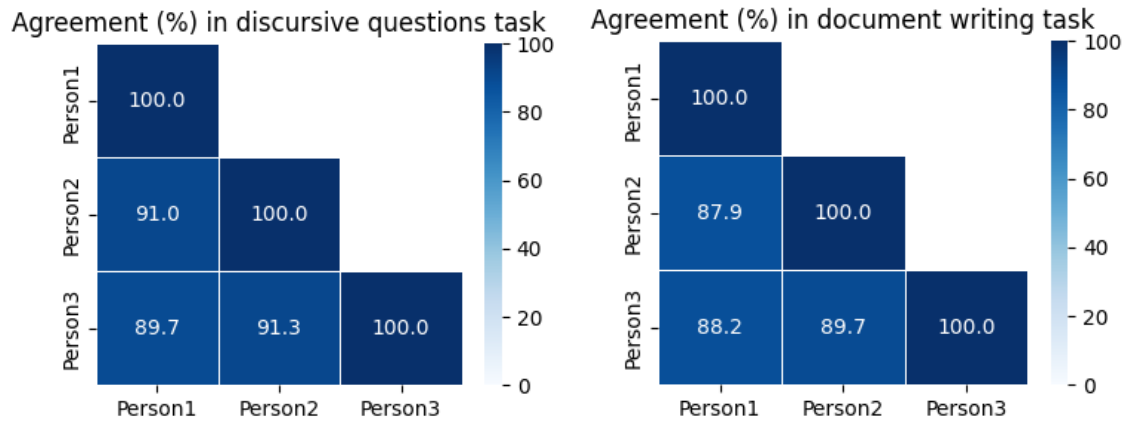


Figure 2: Agreement matrix among human evaluators in two different tasks

We then calculated Cohen’s Kappa between each model’s evaluations and the golden labels. **Figure 3** presents the results: the best-performing evaluator model, gpt4o, achieved a mean Cohen’s Kappa of 79.4 and an average accuracy of 90.9. According to the Landis and Koch (1977) scale, this value is at the upper bound of the “substantial agreement” range (0.61–0.80), and just short of the “almost perfect” category (above 0.80), which further supports its reliability as an evaluator model in this context.

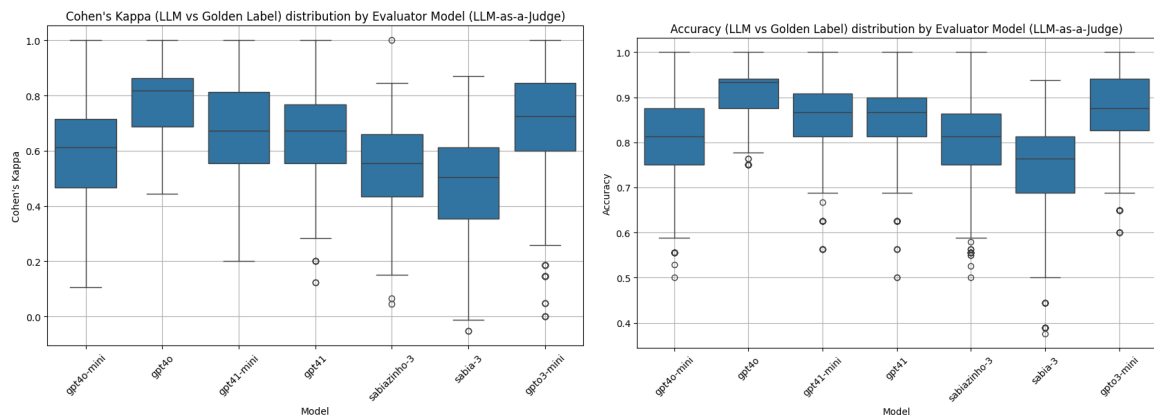


Figure 3: Cohen’s Kappa and Accuracy distribution by Evaluator Model in essay-style questions

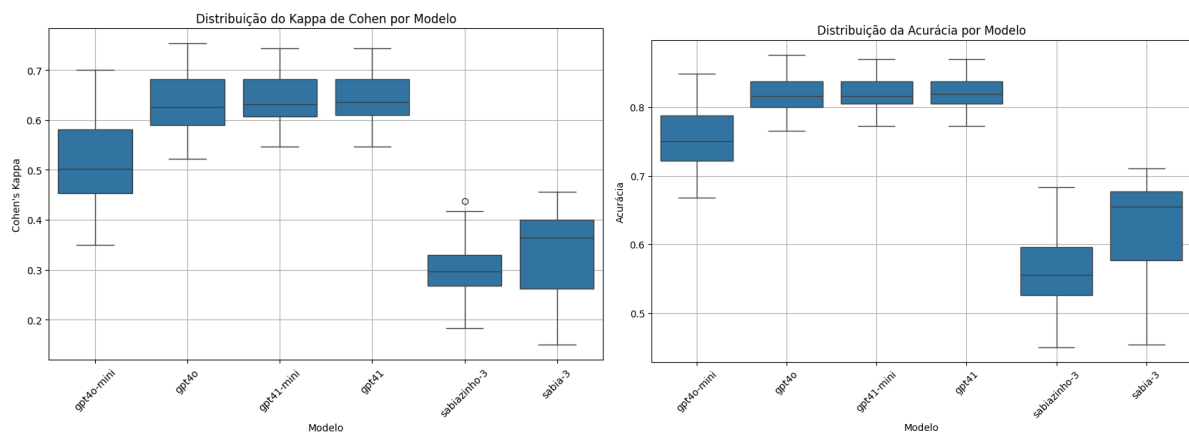


Figure 4: Cohen’s Kappa and Accuracy distribution by Evaluator Model in document writing the models performed worse on the legal drafting task than on the essay questions. The best model’s average Cohen’s Kappa was 64.3—within Landis and Koch’s “substantial agreement” range, but lower than in the essay task. We hypothesize this task is harder, supported by lower inter-rater agreement among humans, as shown by Fleiss’ Kappa and the agreement matrix (**Figure 2**).

3. Results

We applied the benchmark using the best performing model to compare two models gpt4o-mini, (OpenAI) and Sabiazinho-3 (Maritaca.IA). The objective of the test is to understand the strengths and weaknesses of the models, as well as to identify gaps that can be addressed in future versions.

Table 1

Benchmark results

Model	Area	Multiple Choice	Document Choice	Document Writing	Discursive
Gpt-4o-mini	all	64.7	76.19	62.29	58.6
Sabiazinho3	all	63.0	80.95	62.72	62.2

Although the GPT-4o-mini model achieved a slightly higher score on multiple-choice questions, the Sabiazinho-3 model performed better in legal document identification, legal document writing, and answering discursive questions. We can break down this by task and law area to see more details.

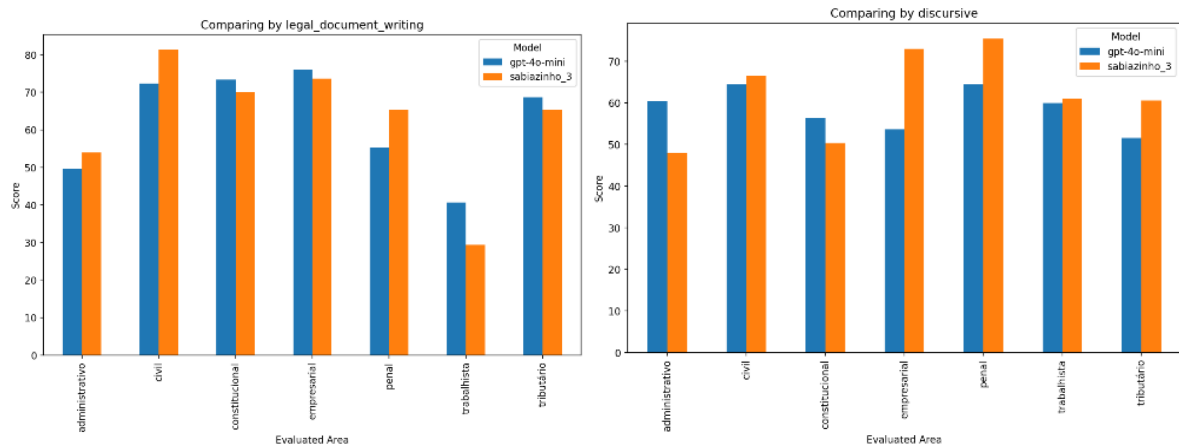


Figure 4: Assessing model’s performance in seven legal subfields

4. Conclusion

Our findings from the human alignment experiment support the existing literature on LLM-as-Judge: there is a high alignment with human evaluations and strong potential for using this technique as an assessment method for open-ended questions. The methodology developed advances the evaluation of LLMs in the Brazilian legal domain. Previously, assessments were limited to multiple-choice tests, which are insufficient as they do not reflect the tasks that legal professionals perform in their daily practice. By creating a benchmark that combines the procedure that qualifies human lawyers to practice law with the models’ ability to evaluate compliance with criteria defined in an official grading rubric, we enable a more in-depth scrutiny of LLMs used by law students, legal professionals, and the public. Our method has the potential to facilitate the development of LLMs for the Brazilian legal domain by allowing better scrutiny, which could lead to improved models for students learning law, professionals practicing it, and the public seeking clarification on whether a particular conduct constitutes a crime.

Declaration on Generative AI

During the preparation of this work, the authors used GPT4o to: Grammar and spelling check. After using these tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

- [1] Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2024. Sabia-3 Technical Report. arXiv preprint arXiv:2410.12049v2 [cs.CL] (Nov. 29, 2024). Available at: <https://arxiv.org/abs/2410.12049>.
- [2] Roseval Malaquias Junior, Ramon Pires, Roseli Romero, and Rodrigo Nogueira. 2024. Juru: Legal Brazilian Large Language Model from Reputable Sources. arXiv:2403.18140 [cs.CL] (Mar. 26, 2024). Available at: <https://arxiv.org/abs/2403.18140>.
- [3] Nicholas Kluge Corrêa, Aniket Sen, Sophia Falk, and Shiza Fatimah. 2024. Tucano: Advancing Neural Text Genera-tion for Portuguese. arXiv:2411.07854 [cs.CL] (Nov. 12, 2024). Available at: <https://arxiv.org/abs/2411.07854>.
- [4] Thales Sales Almeida, Hugo Abonizio, Rodrigo Nogueira, and Ramon Pires. 2024. Sabiá-2: A New Generation of Portuguese Large Language Models. arXiv:2403.09887 [cs.CL] (Mar. 26, 2024). Available at: <https://arxiv.org/abs/2403.09887>.
- [5] Pedro Delfino, Bruno Cuconato, Edward Hermann Haeusler, and Alexandre Rademaker. 2017. Passing the Brazilian OAB Exam: data preparation and some experiments. arXiv:1712.05128v1 [cs.CL] (Dec. 14, 2017). Available at: <https://arxiv.org/abs/1712.05128v1>.
- [6] Evidently AI Team. 2025. LLM-as-a-Judge: A Complete Guide to Using LLMs for Evaluations. Evidently AI (Jan. 9, 2025). Available at: <https://www.evidentlyai.com/llm-guide/llm-as-a-judge>.
- [7] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations? arXiv:2305.01937 [cs.CL] (May 3, 2023). Available at: <https://arxiv.org/abs/2305.01937>.
- [8] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594 [cs.CL] (Jan. 9, 2025). Available at: <https://arxiv.org/abs/2411.15594>.
- [9] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] (June 9, 2023). Available at: <https://arxiv.org/abs/2306.05685>.
- [10] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. arXiv:2303.04048 [cs.CL] (Mar. 7, 2023). Available at: <https://arxiv.org/abs/2303.04048>.
- [11] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods. arXiv:2412.05579v2 [cs.CL] (Dec. 10, 2024). Available at: <https://arxiv.org/abs/2412.05579v2>.
- [12] Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hup-kes. 2025. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. arXiv:2406.12624 [cs.CL] (Jan. 21, 2025). Available at: <https://arxiv.org/abs/2406.12624>.
- [13] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 76, 5 (1971), 378–382. <https://doi.org/10.1037/h0031619>.
- [14] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20, 1 (1960), 37–46. <https://doi.org/10.1177/001316446002000104>.
- [15] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33, 1 (1977), 159–174. <https://doi.org/10.2307/2529310>.