OABench-Consumidor: Um benchmark para avaliação de modelos de linguagem em Direito do Consumidor

Lucas B. Bulcão Mota[1], **Aline Athaydes**[1], **Babacar Mane**[1], **Daniela Barreiro Claro**[1]

[1]FORMAS Research Center on Data and Natural Language
Institute of Computing – Federal University of Bahia (UFBA)
Av. Milton Santos, s/n - Campus de Ondina – 40.170-110 – Salvador – BA – Brazil

{lucasbulcao,alineathaydes,babacarm,dclaro}@ufba.br

Abstract. *Evaluating language models in specific domains, such as Consumer Law, is challenging due to the lack of standardized resources in Brazilian Portuguese. To address this, we introduce OABench-Consumidor, an evaluation benchmark composed exclusively of Consumer Law questions from the Unified Bar Examination (OAB) covering editions from 2010.1 to 2025.2. This paper describes the data collection and structuring process. Furthermore, we demonstrate the benchmark's utility by evaluating the Qwen3-8B foundation model and a fine-tuned version. The specialized model achieved 64.84% accuracy (59/91), outperforming the baseline's 51.65% (47/91). The dataset provides a specialized resource for assessing Artificial Intelligence systems in this legal subarea.*

Resumo. *Avaliar modelos de linguagem em domínios específicos, como o Direito do Consumidor, é um desafio devido à escassez de recursos padronizados em português brasileiro. Para resolver esse problema, introduzimos o OABench-Consumidor, um benchmark composto exclusivamente por questões de Direito do Consumidor da primeira fase do Exame de Ordem Unificado (OAB), cobrindo as edições de 2010.1 a 2025.2. Este artigo descreve a coleta e a estruturação dos dados. Além disso, a utilidade do benchmark é demonstrada através da avaliação do modelo fundacional Qwen3-8B e de uma versão ajustada. O modelo especializado obteve 64,84% de acurácia (59/91), superando os 51,65% (47/91) do modelo base. O benchmark oferece um recurso direto para avaliar sistemas de Inteligência Artificial nessa subárea do Direito.*

1. Introdução

Nos últimos anos, os Grandes Modelos de Linguagem (LLMs) apresentaram avanços significativos em diversas tarefas de Processamento de Linguagem Natural (PLN). Apesar disso, avaliar o desempenho desses modelos no setor jurídico continua sendo um desafio. No Brasil, o desenvolvimento de sistemas de Inteligência Artificial enfrenta a falta de bases de dados e *benchmarks* em português.

O desenvolvimento de sistemas jurídicos confiáveis é essencial em áreas de alto impacto social. O Direito do Consumidor, por exemplo, ocupa uma posição central no judiciário brasileiro. De acordo com [ConJur 2023], 25% de todas as ações distribuídas nas justiças estadual e federal entre 2018 e 2022 tratavam de demandas consumeristas.

Apesar dessa relevância prática, faltam recursos de IA voltados para a área. A ausência de conjuntos de dados dificulta a avaliação da capacidade dos modelos em interpretar normas e resolver casos práticos regidos pelo Código de Defesa do Consumidor (CDC)[Brasil 1990]. Para avaliar esse conhecimento jurídico, uma das referências mais confiáveis no Brasil é o Exame de Ordem Unificado (OAB). A primeira fase do exame utiliza questões de múltipla escolha criadas justamente para testar o conhecimento legal dos candidatos.

A fim de mitigar essa escassez de recursos em PLN, este artigo introduz o **OABench-Consumidor**, um *benchmark* composto apenas por questões de Direito do Consumidor da OAB, abrangendo as edições de 2010.1 a 2025.2. Além de detalhar a construção da base, o trabalho apresenta um experimento comparando o desempenho de um modelo fundacional com uma versão ajustada (*fine-tuned*) para atestar a sensibilidade do conjunto de dados em avaliações de domínio.

2. Trabalhos Relacionados

No Brasil, trabalhos recentes têm buscado reduzir a escassez de datasets especializados para o Direito do Consumidor. Athaydes et al. [Athaydes et al. 2024] propuseram uma metodologia para a construção de um dataset sintético de Perguntas, Respostas e Contextos a partir dos artigos do Código de Defesa do Consumidor (CDC)[Brasil 1990], Súmulas e Acórdãos extraídos do Superior Tribunal de Justiça (STJ). Em continuidade, Athaydes et al. [Athaydes et al. 2025] ampliaram essa proposta ao apresentar uma segunda versão do dataset (V2) e uma avaliação comparativa em relação à versão inicial (V1), discutindo critérios de qualidade do recurso gerado. O modelo ajustado com a V2 é retomado neste artigo na etapa experimental.

Complementando os datasets voltados à adaptação de modelos, benchmarks jurídicos também têm sido propostos para avaliar seu desempenho em tarefas específicas. Exames padronizados de admissão à advocacia, como o *Uniform Bar Examination* (UBE), têm sido usados para testar a capacidade de modelos em interpretar enunciados legais e selecionar respostas juridicamente adequadas. No Brasil, Pires et al. [Pires et al. 2026], da Maritaca AI, propuseram uma metodologia para avaliar LLMs no Exame da OAB, com foco nas questões dissertativas da segunda fase. Este artigo segue a direção de usar a OAB como fonte de avaliação, mas concentra-se nas questões objetivas da primeira fase, com recorte específico em Direito do Consumidor.

No Hugging Face [Hugging Face], também é possível encontrar repositórios com questões da primeira fase da OAB. Durante a inspeção desses repositórios públicos, observou-se ausência de recorte temático confiável para Direito do Consumidor, o que dificulta a avaliação de desempenho em áreas específicas do Direito. O OABench-Consumidor endereça essa limitação ao estruturar um benchmark voltado exclusivamente a questões objetivas de Direito do Consumidor, permitindo avaliar modelos de linguagem em um recorte jurídico mais controlado e diretamente associado ao CDC.

3. Metodologia

Esta seção descreve o processo de construção do *benchmark* OABench-Consumidor, desde a coleta até a organização final dos dados para avaliação automática de modelos de linguagem.

3.1. Coleta e Estruturação de Dados (*Web Scraping*)

O conjunto de dados foi construído a partir de questões da primeira fase do Exame de Ordem Unificado, considerando as edições de 2010.1 a 2025.2. As questões foram extraídas da plataforma *Jurisway*[1], que disponibiliza um histórico de provas da OAB organizado por edição.

A coleta foi realizada por meio de um *script* em Python, utilizando as bibliotecas `requests` e `BeautifulSoup` para acessar as páginas, interpretar o conteúdo HTML e extrair as informações necessárias. O processo seguiu três etapas principais:

1. **Mapeamento das provas:** Inicialmente, o *script* acessa a página de concursos da plataforma, identifica os *links* correspondentes ao Exame de Ordem Unificado e extrai metadados como título, edição e ano da prova.

[1] `https://www.jurisway.org.br/concursos/provas_oab.asp`

2. **Filtragem temática:** Em cada prova identificada, o *script* localiza a seção correspondente à disciplina "Direito do Consumidor" e coleta os *links* das questões associadas a essa área. Para reduzir o risco de sobrecarga no servidor, foi adotado um intervalo de 0,5 segundo entre as requisições.
3. **Extração do conteúdo:** Para cada questão selecionada, foram extraídos o enunciado e as alternativas de múltipla escolha (A, B, C e D). Como o gabarito é disponibilizado em uma página separada no *Jurisway*, o *script* também acessa o *link* de resolução, identificado como "Apenas ver Gabarito", para recuperar a alternativa correta.

3.2. Consolidação e Exportação

Após a extração, os dados foram consolidados em formato tabular utilizando a biblioteca `pandas`. O *benchmark* foi estruturado com as seguintes colunas: *Título*, *Edição*, *Data*, *Pergunta*, *Alternativas* e *Gabarito*. Por fim, o *benchmark* foi exportado para o Hugging Face[2]

4. Experimentos e Resultados

Esta seção apresenta a composição final do OABench-Consumidor e os resultados da avaliação empírica. Inicialmente, são descritos os modelos de linguagem selecionados e o *prompt* utilizado. Na sequência, detalha-se a caracterização geral do *benchmark*. Por fim, compara-se o desempenho do modelo Qwen3-8B, em sua versão base e em sua versão ajustada, na resolução das questões.

4.1. Avaliação de Modelos

Para demonstrar o uso do OABench-Consumidor como instrumento de avaliação, foi conduzido um experimento comparando duas versões do modelo Qwen3-8B:

- **Qwen3-8B (Base):** modelo fundacional avaliado sem ajuste específico para o domínio jurídico.
- **Qwen3-8B (*Fine-Tuned*):** versão do mesmo modelo ajustada para o Direito do Consumidor [Athaydes et al. 2025]. O modelo foi ajustado com um conjunto sintético de perguntas e respostas baseado em artigos do CDC, Súmulas e Acórdãos relacionados à área.

Para padronizar a avaliação, ambos os modelos receberam a mesma instrução, apresentada na Tabela 1, solicitando apenas a letra correspondente à alternativa correta.

Table 1. Estrutura do *prompt* utilizado na avaliação dos modelos.

Prompt de Avaliação
Sua tarefa é, dada a questão de múltipla escolha abaixo, assinale a letra correspondente à resposta correta, levando em consideração o Código de Defesa do Consumidor.
Pergunta: {pergunta}
{alternativas_formatadas}
Escolha a resposta correta replicando APENAS a letra correspondente à resposta correta (ex: "A").
RETORNE APENAS A ALTERNATIVA CORRETA, SEM MAIS NENHUM COMENTÁRIO.

4.2. Caracterização do *benchmark*

A coleta resultou no *benchmark* OABench-Consumidor, composto por 91 questões válidas e formatadas, abrangendo as edições de 2010.1 a 2025.2. A Tabela 2 apresenta as estatísticas gerais do conjunto. Os textos possuem em média 105 *tokens* por enunciado e 120 *tokens* nas alternativas, considerando uma palavra como um token.

[2]Link omitido para garantir o anonimato.

A distribuição dos gabaritos é apresentada na Tabela 3. As alternativas corretas aparecem de forma relativamente equilibrada, variando de 20,9% para a letra B a 28,6% para a letra C. Esse resultado reduz a possibilidade de estratégias baseadas apenas na frequência das alternativas e reforça a adequação do conjunto para avaliação objetiva de modelos.

Table 2. Estatísticas gerais do *benchmark*.

Métrica	Valor
Total de Questões	91
Média de tokens (Enunciado)	104,68
Média de tokens (Alternativas)	120,43

Table 3. Distribuição dos gabaritos no *benchmark*.

Alternativa	Frequência	Percentual
A	21	23,1%
B	19	20,9%
C	26	28,6%
D	25	27,5%

4.3. Comparação dos Modelos

Após a caracterização do conjunto, o *benchmark* OABench-Consumidor foi utilizado para comparar o desempenho dos modelos selecionados. A Tabela 4 apresenta o número de acertos e a acurácia obtida por cada modelo nas 91 questões do *benchmark*.

Table 4. Desempenho dos modelos no OABench-Consumidor.

Modelo	Acertos	Acurácia
Qwen3-8B (Base)	47 de 91	51,65%
Qwen3-8B (*Fine-Tuned*)	59 de 91	64,84%

O modelo base acertou 47 questões (51,65%). O modelo submetido ao ajuste fino com perguntas e respostas sobre os artigos do CDC e jurisprudências alcançou 59 acertos (64,84%). Esse aumento demonstra que a base construída é sensível à adaptação de domínio e pode ser útil como uma ferramenta precisa de avaliação.

Além da diferença na acurácia, também foram observadas diferenças qualitativas no comportamento dos modelos. Durante a execução, o modelo base apresentou tendência a descumprir as restrições do *prompt*, gerando respostas longas ou caracteres desconexos em vez de retornar apenas a alternativa escolhida. Foram realizados testes limitando a quantidade de *tokens* de saída para tentar forçar respostas mais curtas, mas, mesmo assim, o modelo base muitas vezes manteve a tendência de descumprir as restrições impostas. Em contrapartida, o modelo ajustado (*fine-tuned*) demonstrou forte alinhamento à instrução, respondendo de forma objetiva e retornando estritamente apenas a letra da alternativa solicitada.

5. Conclusão

A avaliação de modelos de linguagem no Direito brasileiro exige recursos padronizados e atualizados. Para suprir essa falta, este artigo apresentou o OABench-Consumidor, um *benchmark* focado em Direito do Consumidor, contendo 91 questões de edições históricas da OAB.

A aplicação prática do conjunto validou sua utilidade. A avaliação demonstrou que o modelo Qwen3-8B ajustado para o contexto do CDC superou a versão base, elevando a acurácia de 51,65% para 64,84%. Esses resultados evidenciam que o *benchmark* consegue capturar de forma objetiva a evolução do conhecimento jurídico dos modelos avaliados.

Agradecimentos

Os autores agradecem o financiamento parcial através da bolsa UFBA PIBIC 2025/2026, do apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, FAPESB CCE 0022/2023, FAPESB TIC 002/2015. Este trabalho também conta com o apoio do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILD-IAR) (processo 408490/2024-1). Gostaríamos de agradecer ao Escavador pela parceria neste trabalho.

Referências

Athaydes, A., Bulcao, L., Sacramento, C., Mane, B., Claro, D. B., Souza, M., and Pita, R. (2024). Brazilian consumer protection code: a methodology for a dataset to question-answer (QA) models. In Claro, D. B. and Pagano, A., editors, *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 522–529, Belém do Pará, Brazil. Association for Computational Linguistics.

Athaydes, A., Mota, L., Neto, F. M., Silva, S., Mane, B., Claro, D., Souza, M., and Lisboa, A. (2025). Towards a corpus methodology for llms in the legal domain. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 272–282, Porto Alegre, RS, Brasil. SBC.

Brasil (1990). Lei nº 8.078, de 11 de setembro de 1990. Dispõe sobre a proteção do consumidor e dá outras providências. Diário Oficial da União, Brasília, DF.

ConJur (2023). De cada 4 ações nas justiças estadual e federal, uma é de consumo. `https://www.conjur.com.br/2023-out-11/cada-acoes-justicas-estadual-federal-consumo/`. Acesso em: 21 jun. 2025.

Hugging Face. Hugging face: The ai community building the future. `https://huggingface.co`. Acesso em: março 2026.

Pires, R., Malaquias Junior, R., and Nogueira, R. (2026). Automatic legal writing evaluation of llms. In *Proceedings of the Twentieth International Conference on Artificial Intelligence and Law*, ICAIL '25, pages 420–424, New York, NY, USA. Association for Computing Machinery.