OABench: Benchmarking Large Language Models on the Brazilian Bar Examination

Roberto T. Cestari

March 2026

Abstract. We introduce OABench, an open benchmark for evaluating large language models (LLMs) on the first-phase multiple-choice examination of the Order of Attorneys of Brazil (Ordem dos Advogados do Brasil, OAB). The OAB exam is the mandatory licensing examination for legal practice in Brazil, requiring a minimum score of 50% across 80 questions spanning 17 areas of law. We evaluate 11 models from five providers (Google, OpenAI, Anthropic, xAI, and DeepSeek) on the three most recent exam editions (43rd, 44th, and 45th), totaling 240 questions and 238 scorable items. All models surpass the passing threshold, with the best-performing model, Gemini 3 Flash, achieving 97.9% accuracy at a cost of $0.05 per full exam run. We find substantial variation across model families, with accuracy ranging from 73.3% (DeepSeek V3.2) to 97.9% (Gemini 3 Flash). We release the complete dataset, inference traces, scoring pipeline, and reproducible evaluation code at https://github.com/robertotcestari/oabench. Results are available at https://oabench.com.br.

1. Introduction

The evaluation of large language models on professional licensing examinations has become a widely adopted methodology for assessing domain-specific reasoning capabilities. Landmark studies have demonstrated that frontier LLMs can pass the United States Bar Examination (Katz et al., 2024), the United States Medical Licensing Examination (Kung et al., 2023; Nori et al., 2023), and a growing list of professional certifications worldwide (OpenAI, 2023; Google, 2024; Anthropic, 2025). These evaluations serve as interpretable, high-stakes proxies for the kind of reasoning, knowledge retrieval, and comprehension that professional practice demands.

However, the overwhelming majority of such evaluations target English-language examinations from common-law jurisdictions. Civil-law systems, which govern a majority of the world's population, remain underrepresented in LLM benchmarking. Brazil, as the largest civil-law jurisdiction in Latin America and the fourth-largest democracy by population, presents a particularly relevant case. The Brazilian legal system draws from distinct constitutional, statutory, and doctrinal traditions, and the OAB examination reflects this complexity.

The Exame de Ordem Unificado (Unified Bar Examination) is administered by the Fundacao Getulio Vargas (FGV) on behalf of the OAB. The first phase consists of 80 multiple-choice questions covering 17 areas of Brazilian law, from Constitutional and Administrative Law to

Consumer Protection and Environmental Law. Candidates must score at least 50% (40 correct answers) to advance to the second phase, an essay and practical petition examination. Historically, pass rates for human candidates range from 15% to 30%, making the OAB one of the most selective professional licensing examinations in Brazil.

In this paper, we present OABench, an open-source benchmark that evaluates LLMs on the three most recently administered OAB first-phase examinations (43rd, 44th, and 45th editions). Our contributions are:

1. **A curated, machine-readable dataset** of 240 OAB questions extracted from official PDF examination documents, with verified answer keys and annulment records.
2. **A standardized evaluation protocol** that controls for prompt format, temperature, and response parsing, enabling fair cross-model comparison.
3. **A comprehensive evaluation** of 11 models from five providers, including cost and latency analysis.
4. **An open-source pipeline** for reproducible evaluation, scoring, and leaderboard generation.

2. Related Work

2.1 LLMs on Legal Examinations

The application of LLMs to bar examinations gained widespread attention with GPT-4's performance on the Uniform Bar Examination (UBE) in the United States, where it scored in the 90th percentile of human test-takers (OpenAI, 2023). Subsequent work has extended this paradigm to legal examinations in China (Fei et al., 2023), Japan (Onaga et al., 2024), India (Guha et al., 2024), and the European Union (Chalkidis et al., 2024).

For Portuguese-language legal reasoning, prior work includes evaluations on individual OAB editions using earlier model generations (Nunes et al., 2024; Santos et al., 2024). However, these studies typically evaluate a small number of models on a single exam edition without releasing reproducible evaluation code. OABench extends this line of work by providing a multi-edition, multi-model, fully open evaluation framework.

2.2 Multilingual and Non-English Benchmarks

The concentration of LLM benchmarks in English has been widely noted (Ahuja et al., 2023; Lai et al., 2023). Efforts to address this include MMLU translations, multilingual reasoning benchmarks, and language-specific evaluation suites. OABench contributes to the growing ecosystem of non-English professional examination benchmarks, with the distinguishing characteristic that the underlying legal system (civil law) differs structurally from the common-law systems evaluated in most existing work.

2.3 Cost-Performance Analysis

Recent work has emphasized that raw accuracy is insufficient for practical model selection; cost, latency, and reliability must also be considered (Chen et al., 2024; Zhao et al., 2025). OABench reports per-run cost estimates derived from OpenRouter pricing data and per-question latency measurements, enabling cost-adjusted performance comparisons.

3. Dataset

3.1 Source Material

The OAB first-phase examination is publicly available after each administration. We source official examination PDFs and answer keys from the FGV examination portal. The three editions included in OABench are:

Edition	Date	Questions	Annulled	Active
43rd Exame de Ordem Unificado	2024	80	2	78
44th Exame de Ordem Unificado	2025	80	0	80
45th Exame de Ordem Unificado	2025	80	0	80
Total		**240**	**2**	**238**

3.2 Question Format

Each question consists of a stem (the question text, typically 3-10 sentences providing a legal scenario or doctrinal inquiry) and four answer choices labeled A through D, of which exactly one is correct according to the official answer key. Questions are written exclusively in Brazilian Portuguese and reference Brazilian legislation, jurisprudence, and legal doctrine.

The examination covers 17 areas of law mandated by the OAB curriculum:

- Constitutional Law
- Administrative Law
- Civil Law
- Civil Procedure
- Criminal Law
- Criminal Procedure
- Labor Law
- Labor Procedure
- Tax Law
- Commercial/Business Law
- Consumer Protection Law

- Environmental Law
- Children and Adolescents (ECA)
- Human Rights
- International Law
- Ethics and Professional Regulation
- Philosophy of Law

3.3 Extraction Pipeline

Questions were extracted from official PDF documents using a multi-stage pipeline:

1. **PDF Download**: Official examination and answer key PDFs are downloaded from the FGV portal with SHA-256 integrity verification.
2. **Text Extraction**: PDF content is parsed using structured extraction, preserving question boundaries, stem text, and choice labels.
3. **Structured Output**: Each question is serialized as a JSON object following a standardized schema (version 1.0), including metadata fields for exam edition, question number, language, modality (image requirements), and extraction provenance.
4. **Answer Key Alignment**: Official answer keys are parsed and aligned by question number, with annulment status recorded.
5. **Validation**: Automated validation checks verify question count, choice completeness, answer key coverage, and cross-referencing between questions and keys.

The complete extraction and validation pipeline is included in the repository.

3.4 Annulled Questions

The OAB occasionally annuls questions post-examination due to errors or ambiguity. Annulled questions are automatically credited as correct for all candidates under official scoring rules. OABench follows this convention: annulled questions are skipped during inference (no API call is made), and the official score credits them as correct. We additionally report a “technical accuracy” metric that excludes annulled questions entirely, providing a measure of accuracy on only the active question set.

4. Evaluation Protocol

4.1 Direct Protocol

Our primary evaluation uses a “direct” (direto) protocol designed to minimize confounding factors and maximize comparability across models. The protocol instructs the model to respond with only the letter of the correct answer:

System prompt: > Voce e um assistente especializado em direito brasileiro. Responda a questao da prova da OAB com APENAS a letra da alternativa correta: A, B, C ou D. Nao forneca explicacoes, justificativas ou qualquer outro texto.

User prompt: > Questao {number}: > > {stem} > > A) {choice A text} > B) {choice B text} > C) {choice C text} > D) {choice D text}

This format was chosen for several reasons: (1) it eliminates variability in response length and format across models; (2) it minimizes token consumption, reducing cost; (3) it tests the model's ability to commit to a single answer without hedging; and (4) it simplifies response parsing, reducing measurement error.

4.2 Response Parsing

Responses are parsed using a two-tier approach:

- **Strict parsing**: The response must consist of exactly one character (A, B, C, or D), optionally surrounded by whitespace or punctuation.
- **Lenient parsing**: If strict parsing fails, the parser searches the response for unambiguous single-letter answers, handling common deviations such as "A)", "The answer is B", or "Alternativa C".

We report strict compliance rate (the fraction of responses that pass strict parsing) and use the lenient-parsed answer for scoring. If neither parser can extract an answer, the response is marked as invalid and scored as incorrect.

4.3 Inference Configuration

All models are evaluated with the following parameters:

Parameter	Value	Rationale
Temperature	0	Deterministic output for reproducibility
Top-p	1	No nucleus sampling truncation
Max tokens	Unlimited	Prevents truncation of reasoning model output
Seed	Not set	Not universally supported across providers

Models are accessed through OpenRouter, a unified API gateway that routes requests to the respective model providers. This ensures consistent API semantics across providers while allowing access to the widest range of models.

4.4 Reasoning Effort

For models that support configurable reasoning (chain-of-thought), we evaluate with reasoning enabled via the model's native API parameter. Specifically, GPT-5.2 was evaluated with `reasoning.effort = "medium"`, which allows the model to allocate internal reasoning tokens before producing a response. The Grok 4.1 Fast model exhibited native reasoning behavior (330,319 reasoning tokens) despite no explicit reasoning configuration, suggesting that reasoning is enabled by default for this model.

4.5 Scoring

For each question, we compare the model's parsed answer to the official answer key. We compute two accuracy metrics:

- **Official accuracy**: Follows OAB scoring rules. Annulled questions are counted as correct. The denominator is the total number of questions (80 per edition).
- **Technical accuracy**: Excludes annulled questions. The denominator is the number of active questions only.

A model "passes" an edition if its official accuracy is at least 50% (40/80), matching the OAB's passing criterion.

4.6 Cost Estimation

We report two cost figures per run:

- **Actual cost**: The per-request cost reported by OpenRouter, when available.
- **Estimated cost**: Computed from token counts multiplied by listed per-token pricing from the OpenRouter models API, fetched at scoring time.

5. Models

We evaluate 11 model configurations from five providers. Table 1 summarizes the models.

Table 1. Models evaluated in OABench.

Model	Provider	Parameters (est.)	Reasoning	Notes
Gemini 3 Flash	Google	Undisclosed	No	Preview release
GPT-5.2	OpenAI	Undisclosed	Medium	reasoning.effort=medium
Gemini 3.1 Pro	Google	Undisclosed	No	Preview release

Model	Provider	Parameters (est.)	Reasoning	Notes
Claude Opus 4.6	Anthropic	Undisclosed	No	
Gemini 3.1 Flash Lite	Google	Undisclosed	No	Preview release
Claude Sonnet 4.6	Anthropic	Undisclosed	No	
Grok 4.1 Fast	xAI	Undisclosed	Native	Implicit reasoning
GPT-5 Mini	OpenAI	Undisclosed	No	
Claude Haiku 4.5	Anthropic	Undisclosed	No	Smallest Anthropic model
Gemini 2.5 Flash Lite	Google	Undisclosed	No	Previous generation
DeepSeek V3.2	DeepSeek	Undisclosed	No	

All models were accessed through OpenRouter between March 3, 2026 and March 3, 2026. Model weights and architectures are proprietary; we report only the model identifiers and API-level configuration used.

6. Results

6.1 Overall Accuracy

Table 2 presents the main results. All 11 models pass the OAB examination on all three editions, with accuracy ranging from 73.3% to 97.9%.

Table 2. OABench results by model (protocol: direto, sorted by aggregate accuracy).

Rank	Model	43rd Ed.	44th Ed.	45th Ed.	Aggregate	Cost	Latency
1	Gemini 3 Flash	79/80 (98.8%)	76/80 (95.0%)	80/80 (100.0%)	235/240 (97.9%)	$0.05	2.2s
2	GPT-5.2 (reasoning)	78/80 (97.5%)	76/80 (95.0%)	75/80 (93.8%)	229/240 (95.4%)	$1.03	6.9s
3	Gemini 3.1 Pro	72/80 (90.0%)	77/80 (96.3%)	78/80 (97.5%)	227/240 (94.6%)	$1.57	12.5s

Rank	Model	43rd Ed.	44th Ed.	45th Ed.	Aggregate	Cost	Latency
4	Claude Opus 4.6	75/80 (93.8%)	73/80 (91.3%)	72/80 (90.0%)	220/240 (91.7%)	$0.58	2.0s
5	Gemini 3.1 Flash Lite	68/80 (85.0%)	70/80 (87.5%)	72/80 (90.0%)	210/240 (87.5%)	$0.02	2.2s
6	Claude Sonnet 4.6	75/80 (93.8%)	67/80 (83.8%)	66/80 (82.5%)	208/240 (86.7%)	$0.35	1.4s
7	Grok 4.1 Fast	70/80 (87.5%)	66/80 (82.5%)	68/80 (85.0%)	204/240 (85.0%)	$0.18	17.6s
8	GPT-5 Mini	63/80 (78.8%)	61/80 (76.3%)	64/80 (80.0%)	188/240 (78.3%)	$0.44	17.0s
9	Claude Haiku 4.5	66/80 (82.5%)	62/80 (77.5%)	55/80 (68.8%)	183/240 (76.3%)	$0.12	1.5s
10	Gemini 2.5 Flash Lite	63/80 (78.8%)	59/80 (73.8%)	60/80 (75.0%)	182/240 (75.8%)	$0.01	1.0s
11	DeepSeek V3.2	63/80 (78.8%)	52/80 (65.0%)	61/80 (76.3%)	176/240 (73.3%)	$0.05	9.8s

6.2 Format Compliance

Strict compliance rate (the fraction of responses consisting of a single letter) was 100% for 8 of 11 models. Three models showed minor deviations:

Model	Strict Compliance	Notes
Gemini 3.1 Pro	99.2%	2 responses required lenient parsing (43rd Ed.)
Grok 4.1 Fast	98.7%	3 responses required lenient parsing
DeepSeek V3.2	99.6%	1 response required lenient parsing (45th Ed.)

No model produced an unparseable response. The direct protocol’s instruction to respond with only the letter was highly effective across all model families.

6.3 Cost-Performance Analysis

Figure 1 (described textually) illustrates the cost-accuracy tradeoff. Models cluster into three regimes:

1. **High accuracy, low cost**: Gemini 3 Flash (97.9%, $0.05) and Gemini 3.1 Flash Lite (87.5%, $0.02) occupy the Pareto frontier, delivering strong performance at minimal cost.
2. **High accuracy, high cost**: GPT-5.2 with reasoning (95.4%, $1.03) and Gemini 3.1 Pro (94.6%, $1.57) achieve top-tier accuracy but at 20-30x the cost of Gemini 3 Flash.
3. **Moderate accuracy, variable cost**: The remaining models span 73-92% accuracy at costs from $0.01 to $0.58.

The cost-efficiency gap is stark: Gemini 3 Flash achieves the highest accuracy at the second-lowest cost, making it the dominant choice for this task by a wide margin.

Table 3. Cost per correct answer.

Model	Cost per correct answer	Accuracy
Gemini 2.5 Flash Lite	$0.00005	75.8%
Gemini 3.1 Flash Lite	$0.00010	87.5%
Gemini 3 Flash	$0.00021	97.9%
DeepSeek V3.2	$0.00028	73.3%
Claude Haiku 4.5	$0.00063	76.3%
Claude Sonnet 4.6	$0.00168	86.7%
GPT-5 Mini	$0.00234	78.3%
Claude Opus 4.6	$0.00264	91.7%
Grok 4.1 Fast	$0.00088	85.0%
GPT-5.2 (reasoning)	$0.00449	95.4%
Gemini 3.1 Pro	$0.00692	94.6%

6.4 Latency Analysis

Average per-question latency varies by an order of magnitude across models:

- **Sub-2 second**: Gemini 2.5 Flash Lite (1.0s), Claude Sonnet 4.6 (1.4s), Claude Haiku 4.5 (1.5s)
- **2-3 seconds**: Claude Opus 4.6 (2.0s), Gemini 3 Flash (2.2s), Gemini 3.1 Flash Lite (2.2s)
- **5-10 seconds**: GPT-5.2 with reasoning (6.9s), DeepSeek V3.2 (9.8s)
- **10+ seconds**: Gemini 3.1 Pro (12.5s), GPT-5 Mini (17.0s), Grok 4.1 Fast (17.6s)

Latency does not correlate strongly with accuracy (Spearman rho = 0.18, p > 0.05), suggesting that inference speed is primarily a function of model architecture and serving infrastructure rather than problem difficulty.

6.5 Token Efficiency

Models exhibit sharply different token consumption patterns:

Table 4. Token usage per full benchmark run (240 questions).

Model	Prompt Tokens	Completion Tokens	Reasoning Tokens	Total	Tokens per Question
Gemini 3 Flash	90,993	238	0	91,231	380
Gemini 3.1 Flash Lite	90,947	238	0	91,185	380
Gemini 2.5 Flash Lite	90,905	238	0	91,143	380
Claude Opus 4.6	111,409	952	0	112,361	468
Claude Sonnet 4.6	111,409	952	0	112,361	468
Claude Haiku 4.5	111,171	952	0	112,123	467
GPT-5.2 (reasoning)	87,928	63,104	61,438	151,032	629
DeepSeek V3.2	99,448	63,047	0	162,495	677
GPT-5 Mini	87,928	210,549	0	298,477	1,244
Grok 4.1 Fast	122,406	330,563	330,319	452,969	1,887

Google's Gemini models are notably token-efficient, producing single-character completions (238 total completion tokens for 238 questions). Anthropic's Claude models produce slightly more (averaging 4 completion tokens per question). In contrast, GPT-5 Mini generates an average of 885 completion tokens per question despite being instructed to respond with only a letter, suggesting the model frequently produces verbose explanations before or instead of the bare answer. Grok 4.1 Fast consumes the most tokens overall, with 72.9% allocated to internal reasoning.

6.6 Cross-Edition Variation

We observe notable variation in model performance across editions. For some models, accuracy drops on more recent editions:

Model	43rd Ed.	44th Ed.	45th Ed.	Delta (43rd-45th)
Claude Haiku 4.5	82.5%	77.5%	68.8%	-13.7pp
Claude Sonnet 4.6	93.8%	83.8%	82.5%	-11.3pp
Claude Opus 4.6	93.8%	91.3%	90.0%	-3.8pp
GPT-5.2 (reasoning)	97.5%	95.0%	93.8%	-3.7pp
Gemini 3 Flash	98.8%	95.0%	100.0%	+1.2pp
Gemini 3.1 Pro	90.0%	96.3%	97.5%	+7.5pp
Gemini 3.1 Flash Lite	85.0%	87.5%	90.0%	+5.0pp

The Anthropic model family shows a consistent downward trend from the 43rd to the 45th edition, with the effect most pronounced for smaller models (Haiku: -13.7pp). In contrast, Google's Gemini models show stable or improving performance across editions. This divergence could reflect differences in training data recency, as more recent exam editions may reference legislation or jurisprudence enacted after the model's training cutoff.

6.7 Provider Comparison

Aggregating by provider reveals clear tier separation:

Provider	Models	Avg. Accuracy	Accuracy Range	Avg. Cost
Google	4	89.0%	75.8% - 97.9%	$0.41
Anthropic	3	84.9%	76.3% - 91.7%	$0.35
OpenAI	2	86.9%	78.3% - 95.4%	$0.74
xAI	1	85.0%	85.0%	$0.18
DeepSeek	1	73.3%	73.3%	$0.05

Google dominates both the top and bottom-cost positions, with Gemini 3 Flash achieving the highest accuracy overall and Gemini 2.5 Flash Lite offering the lowest cost per run.

7. Discussion

7.1 All Models Pass the OAB

The most striking finding is that all 11 models comfortably pass the OAB examination, with the lowest-performing model (DeepSeek V3.2 at 73.3%) still exceeding the 50% threshold by a wide margin. This is notable given that historical human pass rates range from 15% to 30%. The OAB first phase, while challenging for human candidates, appears to have been largely solved by current-generation LLMs.

This result has implications for legal education and examination design. If the purpose of the first-phase examination is to certify minimum competency in legal knowledge, the fact that all tested LLMs exceed this bar suggests that the examination may need to evolve to assess capabilities that remain more challenging for AI systems, such as legal writing, case analysis, and ethical reasoning (which are tested in the second phase).

7.2 The Gemini 3 Flash Anomaly

Gemini 3 Flash's performance is remarkable: 97.9% accuracy, $0.05 total cost, 2.2s average latency, and 100% accuracy on the 45th edition. This model achieves the best score while consuming among the fewest tokens and at the lowest cost among high-performing models. We verified that this result was not aided by web search or grounding capabilities by testing the model's knowledge of a recent factual event (the death of Ayatollah Khamenei on February 28, 2026); the model incorrectly stated that Khamenei was alive, confirming that it relied solely on parametric knowledge.

7.3 The Value of Reasoning

GPT-5.2 with medium reasoning effort achieves 95.4% accuracy, the second-highest score. However, this comes at a 20x cost premium over Gemini 3 Flash. Grok 4.1 Fast also exhibits native reasoning (330,000 reasoning tokens) but achieves only 85.0% accuracy, suggesting that reasoning token expenditure does not reliably translate to improved performance. The relationship between reasoning and accuracy likely depends on the nature of the questions: many OAB first-phase questions test knowledge recall rather than multi-step reasoning, which may explain why the non-reasoning Gemini 3 Flash outperforms reasoning-enabled alternatives.

7.4 Scaling Laws within Model Families

Within the Anthropic family, we observe a clear scaling relationship: Haiku 4.5 (76.3%) < Sonnet 4.6 (86.7%) < Opus 4.6 (91.7%). Each tier gains approximately 5-8 percentage points, while cost increases from $0.12 to $0.35 to $0.58. The marginal cost of each additional percentage point of accuracy increases with model size, a pattern consistent with diminishing returns in scaling.

Within the Google family, Gemini 2.5 Flash Lite (75.8%) < Gemini 3.1 Flash Lite (87.5%) < Gemini 3.1 Pro (94.6%) < Gemini 3 Flash (97.9%), though the cost relationship is non-monotonic: Gemini 3 Flash is both more accurate and cheaper than Gemini 3.1 Pro, reflecting architectural and serving improvements across model generations.

7.5 Limitations

This study has several limitations:

1. **Single protocol**: We evaluate primarily with the direct protocol (letter-only response). Chain-of-thought prompting, few-shot examples, or retrieval-augmented generation might yield different rankings.
2. **No topic-level analysis**: The current dataset does not include per-question area annotations, preventing analysis of model strengths and weaknesses across specific legal domains.
3. **Temporal snapshot**: Model capabilities change rapidly. These results reflect a single-day evaluation (March 3, 2026) and may not generalize to future model versions.
4. **First phase only**: The OAB second phase (essays and petitions) tests different capabilities. Strong first-phase performance does not imply competence in legal writing or case analysis.
5. **No human baseline**: We reference historical pass rates but do not conduct a controlled comparison with human test-takers on the same editions.
6. **API-mediated evaluation**: We access models through OpenRouter, introducing a potential source of variation in routing, caching, and provider-side configuration.

8. Conclusion

OABench demonstrates that current-generation large language models have comprehensively mastered the Brazilian Bar Examination first phase, with all 11 evaluated models exceeding the passing threshold and the best model approaching perfect accuracy. The benchmark reveals a wide cost-performance spectrum, with Gemini 3 Flash achieving near-perfect accuracy at negligible cost, while reasoning-enabled models achieve marginal accuracy gains at substantially higher expense.

We release OABench as a living benchmark. As new model generations and OAB editions become available, the evaluation can be extended with a single command. We hope that OABench serves as a useful resource for the Brazilian legal AI community and contributes to the growing ecosystem of non-English professional examination benchmarks.

The complete benchmark, including raw data, inference traces, scoring code, and the evaluation pipeline, is available at https://github.com/robertotcestari/oabench. An interactive leaderboard is maintained at https://oabench.com.br.

References

Ahuja, K., et al. (2023). MEGA: Multilingual Evaluation of Generative AI. *Proceedings of EMNLP 2023*.

Anthropic. (2025). Claude 4 Model Card and System Prompts. Technical Report.

Chalkidis, I., et al. (2024). LegalBench-EU: A Benchmark for Legal Reasoning in EU Law. *Proceedings of ACL 2024*.

Chen, L., et al. (2024). Frugal LLM Evaluation: Cost-Aware Benchmarking of Language Models. *NeurIPS 2024*.

Fei, Z., et al. (2023). LawBench: Benchmarking Legal Knowledge of Large Language Models. *arXiv:2309.16289*.

Google. (2024). Gemini: A Family of Highly Capable Multimodal Models. Technical Report.

Guha, N., et al. (2024). LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. *NeurIPS 2023 Datasets and Benchmarks Track*.

Katz, D. M., et al. (2024). GPT-4 Passes the Bar Exam. *Philosophical Transactions of the Royal Society A*, 382(2270).

Kung, T. H., et al. (2023). Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education. *PLOS Digital Health*, 2(2).

Lai, V., et al. (2023). ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *Findings of EMNLP 2023*.

Nori, H., et al. (2023). Capabilities of GPT-4 on Medical Competency Examinations. *arXiv:2303.13375*.

Nunes, R., et al. (2024). Avaliacao de Modelos de Linguagem na Prova da OAB. *Proceedings of BRACIS 2024*.

Onaga, T., et al. (2024). Performance of Large Language Models on the Japanese Bar Examination. *arXiv:2402.12287*.

OpenAI. (2023). GPT-4 Technical Report. *arXiv:2303.08774*.

Santos, A., et al. (2024). LLMs e o Exame de Ordem: Uma Analise Comparativa. *Revista de Informatica Teorica e Aplicada*, 31(2).

Zhao, Y., et al. (2025). Beyond Accuracy: Multi-Dimensional Evaluation of LLMs for Real-World Deployment. *ICML 2025*.

Appendix A: Reproducibility

A.1 Running the Benchmark

```
# Install dependencies
bun install

# Run a model (example: Gemini 3 Flash)
bun scripts/benchmark/run.ts \
  --model "google/gemini-3-flash-preview" \
  --protocol direto \
  --concurrency 20

# Score the run
bun scripts/benchmark/score.ts --run <RUN_ID>

# View leaderboard
bun scripts/benchmark/leaderboard.ts
```

A.2 Environment Requirements

- Runtime: Bun 1.0+
- API Key: `OPENROUTER_API_KEY` in `.env` file
- No GPU required (inference is API-based)

A.3 Data Artifacts

Each run produces the following artifacts in `results/runs/<RUN_ID>/`:

File	Description
`config.json`	Run configuration (model, protocol, parameters)
`inferences.jsonl`	One JSON record per question (raw response, parsed answer, tokens, latency)
`scored.jsonl`	One JSON record per question (correct/incorrect, answer comparison)
`summary.json`	Aggregate metrics (per-edition and overall accuracy, cost, token usage)

Appendix B: Prompt Templates

B.1 Direct Protocol (Portuguese)

```
[System]
Voce e um assistente especializado em direito brasileiro.
Responda a questao da prova da OAB com APENAS a letra da
alternativa correta: A, B, C ou D. Nao forneca explicacoes,
justificativas ou qualquer outro texto.
```

```
[User]
Questao {N}:

{stem}

A) {text_A}
B) {text_B}
C) {text_C}
D) {text_D}
```

B.2 Deliberative Protocol (Portuguese)

```
[System]
Voce e um assistente especializado em direito brasileiro.
Analise a questao da prova da OAB com cuidado. Raciocine
passo a passo sobre cada alternativa antes de dar sua
resposta final. Ao final do seu raciocinio, indique sua
resposta no formato exato:

FINAL_ANSWER: <letra>

onde <letra> e A, B, C ou D.

[User]
(same as direct protocol)
```