Universidade Federal de Sergipe

Departamento de Computação

Tópicos Avançados em Engenharia de Software e Sistemas de Informação I

Atividade Avaliativa 1

Curadoria de Datasets e Inferência Básica com LLMs

Equipe 3: Domínio Jurídico

Reinan Gabriel dos Santos Souza

Fernanda Mirely Barbosa Souza

Éricles dos Santos Cunha

Júlia de souza lima

Mikaela de Andrade Lima

Victor Leonardo Mascarenhas Soares Horta

Abril de 2026

Vídeo demonstrativo:

https://youtu.be/1QMsz2wOlt0

Sumário

1 Introdução

Este documento apresenta os resultados da atividade avaliativa 1 da disciplina Tópicos Avançados em Engenharia de Software e Sistemas de Informação I, ministrada na Universidade Federal de Sergipe (UFS), semestre 2026.1.

A atividade consiste na curadoria de datasets jurídicos e na realização de inferência básica utilizando Modelos de Linguagem (LLMs), com foco em questões do Exame da OAB (Ordem dos Advogados do Brasil). Embora a execução das tarefas de curadoria e análise seja individual, a entrega e consolidação dos resultados são feitas em equipe.

O presente relatório documenta as contribuições da **Equipe 3 (Jurídica)**. Cada membro da equipe ficou responsável por um lote específico de questões, e suas contribuições individuais são detalhadas na Seção 2.

2 Contribuições individuais

Esta seção detalha as contribuições de cada membro da equipe na execução da atividade.

2.1 Reinan Gabriel dos Santos Souza

Toda a implementação do pipeline, desde a curadoria das questões até a geração dos gráficos de avaliação, está disponível no repositório público no GitHub[1]. O repositório segue a abordagem *Docs-as-Code*, com documentação técnica versionada. A partir dele é possível acessar a documentação web e navegar pelos artefatos para visualizar os resultados obtidos.

2.1.1 Escolha e justificativa dos modelos

A seleção dos modelos foi condicionada pelo hardware disponível, especificamente uma GPU NVIDIA GeForce GTX 1050 com apenas 4 GB de VRAM dedicada. Essa restrição limitou a escolha a LLMs compactos, com até aproximadamente 3 bilhões de parâmetros, em versões quantizadas que não ultrapassassem 2 GB de armazenamento.

Foram selecionados três modelos de organizações distintas, todos executados localmente via Ollama, a saber, o **Gemma 2 2B**, o **Llama 3.2 3B** e o **Qwen 2.5 3B**. A escolha priorizou três critérios, compatibilidade com o hardware disponível, diversidade de desenvolvedores para evitar viés arquitetural e suporte ao idioma português, requisito essencial para o domínio jurídico brasileiro.

[1] https://github.com/reinanhs/Topicos_Avancados_2026_1_Equipe_JUD_3_atividade1

2.1.2 Resultados da avaliação exata

A Tabela 1 apresenta o desempenho dos três modelos nas 122 questões objetivas do lote atribuído, avaliados por meio de Acurácia, Precisão, Recall e F1-Score. O **gemma2:2b** obteve a maior acurácia (0,4344), seguido pelo **qwen2.5:3b** (0,4180) e **llama3.2:3b** (0,3607). Nenhum modelo ultrapassou 50%, resultado esperado considerando o tamanho reduzido dos modelos e a complexidade inerente às questões da OAB.

Tabela 1: Avaliação exata do desempenho dos modelos em questões objetivas

Modelo	Acurácia	Precisão	Recall	F1
gemma2:2b	0,4344	0,4322	0,4448	0,4101
llama3.2:3b	0,3607	0,3519	0,3563	0,3473
qwen2.5:3b	0,4180	0,3609	0,3405	0,3337

A Figura 1 permite visualizar essa comparação de forma mais direta. Observa-se que o **gemma2:2b** mantém valores ligeiramente superiores em todas as métricas, enquanto o **llama3.2:3b** fica consistentemente abaixo dos demais. A Figura 2 complementa essa análise com um gráfico radar, no qual o **gemma2:2b** apresenta o polígono mais simétrico e amplo, indicando o perfil mais equilibrado entre os três modelos.

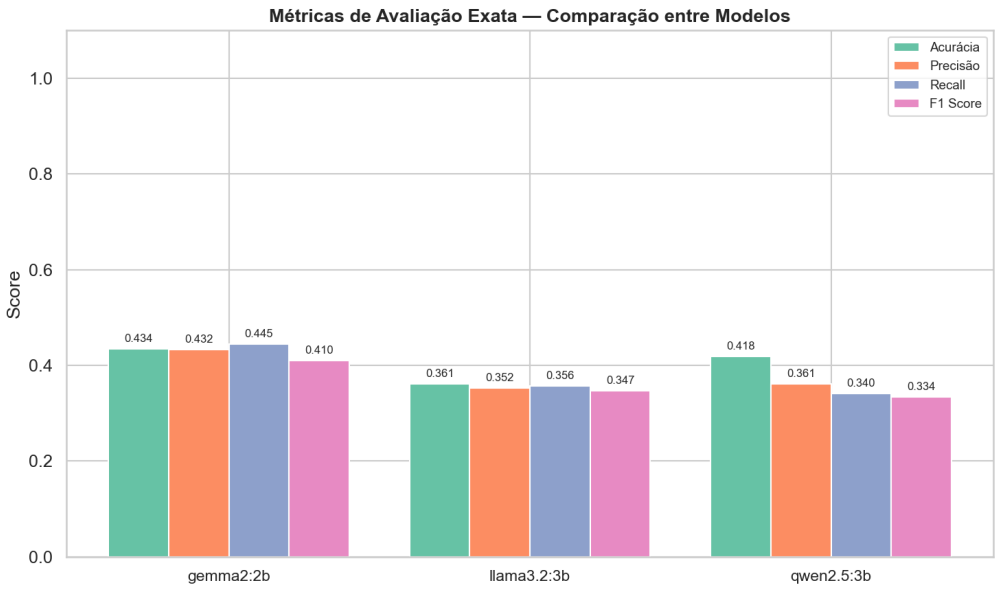

Figura 1: Métricas de avaliação exata na comparação entre modelos

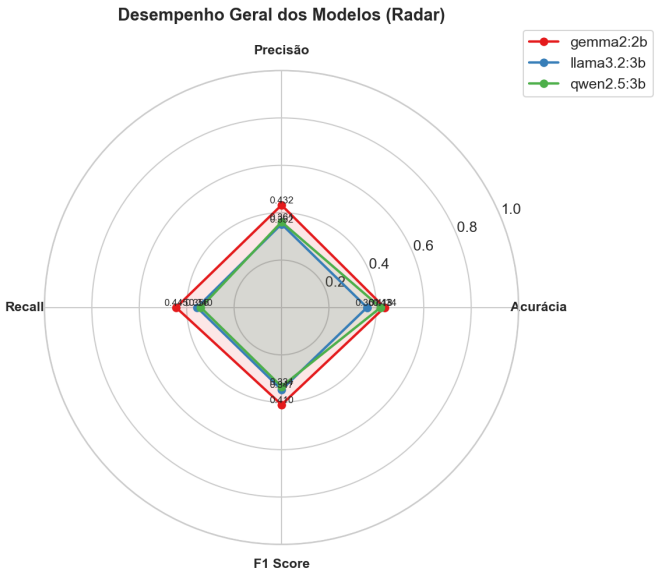

Figura 2: Desempenho geral dos modelos em gráfico radar

2.1.3 Resultados da avaliação cruzada

Para as 12 questões dissertativas, que não possuem gabarito oficial, a avaliação foi conduzida de duas formas: comparação entre as respostas dos próprios modelos e comparação com a guideline de referência fornecida pelo dataset. A Tabela 2 apresenta as métricas de similaridade entre os pares de modelos, enquanto a Tabela 3 mostra a aderência de cada modelo à guideline.

Tabela 2: Avaliação cruzada de similaridade entre pares de modelos

Par de modelos	**BLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore F1**
gemma2:2b vs llama3.2:3b	0,1205	0,4575	0,1789	0,2459	0,7319
gemma2:2b vs qwen2.5:3b	0,1011	0,4741	0,1734	0,2256	0,7394
llama3.2:3b vs qwen2.5:3b	0,1025	0,4473	0,1664	0,2320	0,7323

Tabela 3: Aderência dos modelos à guideline de referência.

Modelo vs Guideline	**BLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore F1**
gemma2:2b vs guideline	0,0216	0,1534	0,0585	0,1126	0,6569
llama3.2:3b vs guideline	0,0215	0,1600	0,0571	0,1113	0,6632
qwen2.5:3b vs guideline	0,0172	0,1260	0,0447	0,0910	0,6513

Os modelos apresentam concordância semântica razoável entre si, com BERTScore F1 acima de 0,73 em todos os pares, apesar da baixa sobreposição lexical (BLEU $\approx$ 0,10–0,12). Já a similaridade com a guideline é consideravelmente menor (BERTScore F1 em

torno de 0,65), indicando que, embora os modelos concordem entre si, suas respostas se distanciam do padrão de referência esperado.

A Figura 3 ilustra essa diferença por meio de um heatmap de similaridade semântica. As células entre modelos apresentam tons mais intensos (maior similaridade), enquanto as comparações com a guideline aparecem em tons mais claros.

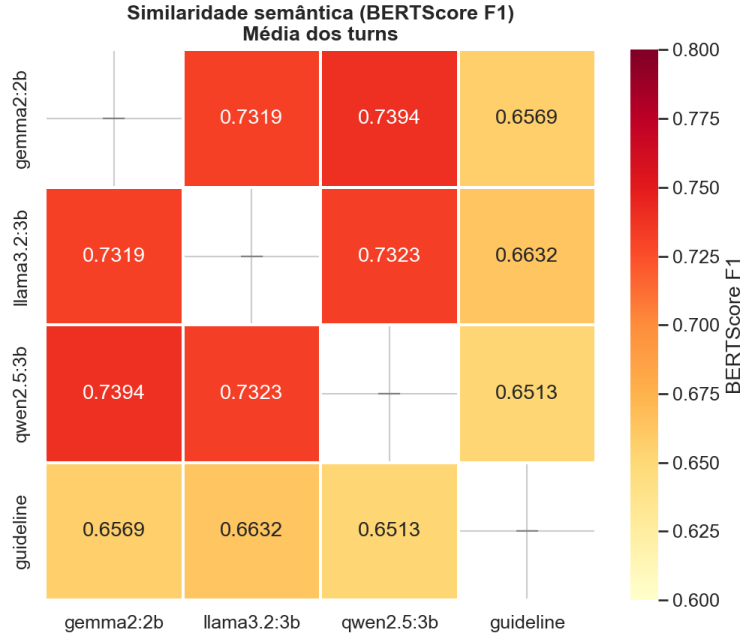

Figura 3: Heatmap de similaridade semântica entre modelos e guideline.

A Figura 4 detalha todas as métricas de similaridade entre os pares de modelos. O BERTScore F1 é consistentemente a métrica mais elevada, enquanto o BLEU permanece baixo, resultado esperado, pois o BERTScore captura a semântica contextual, ao passo que o BLEU depende da correspondência exata de n-gramas.

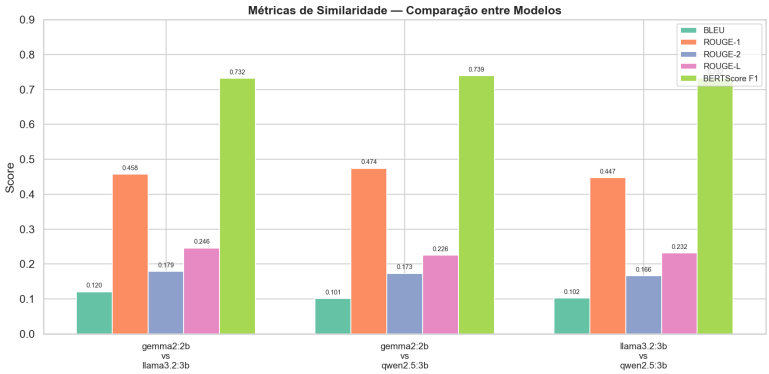

Figura 4: Comparação entre pares de modelos com métricas de similaridade

A Figura 5 apresenta a aderência de cada modelo à guideline. Os valores baixos de BLEU e ROUGE confirmam que os modelos não reproduzem literalmente o texto de referência, embora o BERTScore em torno de 0,65 sugira certa proximidade semântica.

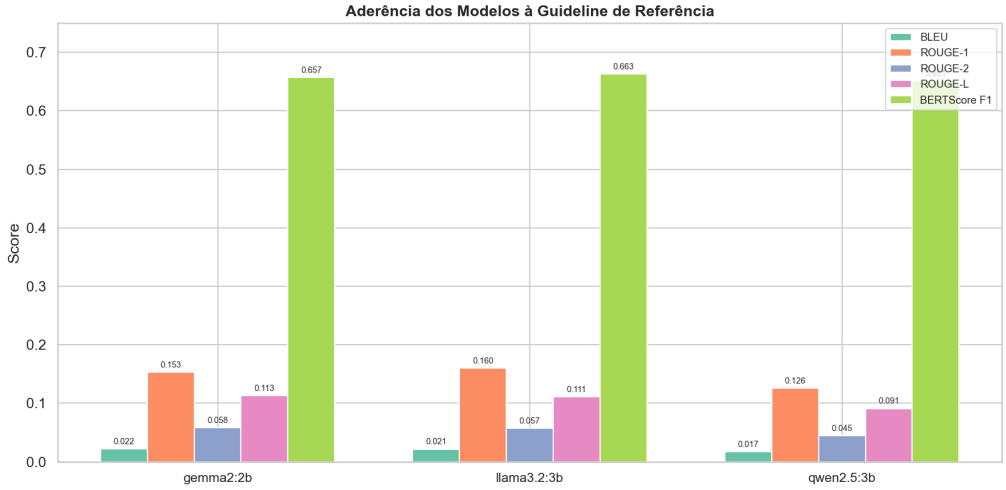

Figura 5: Aderência dos modelos à guideline de referência.

A Figura 6 complementa essa análise com gráficos radar individuais. Todos os modelos apresentam um padrão semelhante: alta concentração no eixo do BERTScore F1 e valores baixos nas demais métricas, reforçando a diferença entre similaridade lexical e semântica.

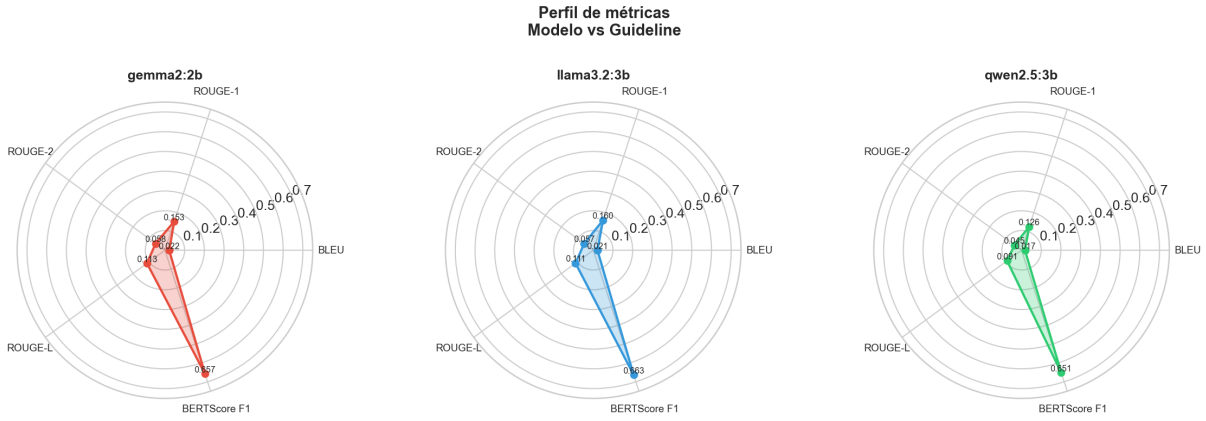

Figura 6: Perfil de métricas por modelo em relação à guideline.

A Figura 7 desagrega o BERTScore F1 por turn. O Turn 1 tende a apresentar scores ligeiramente superiores ao Turn 2 em praticamente todos os pares, o que pode indicar que a segunda parte das questões exige respostas mais específicas, nas quais os modelos divergem mais.

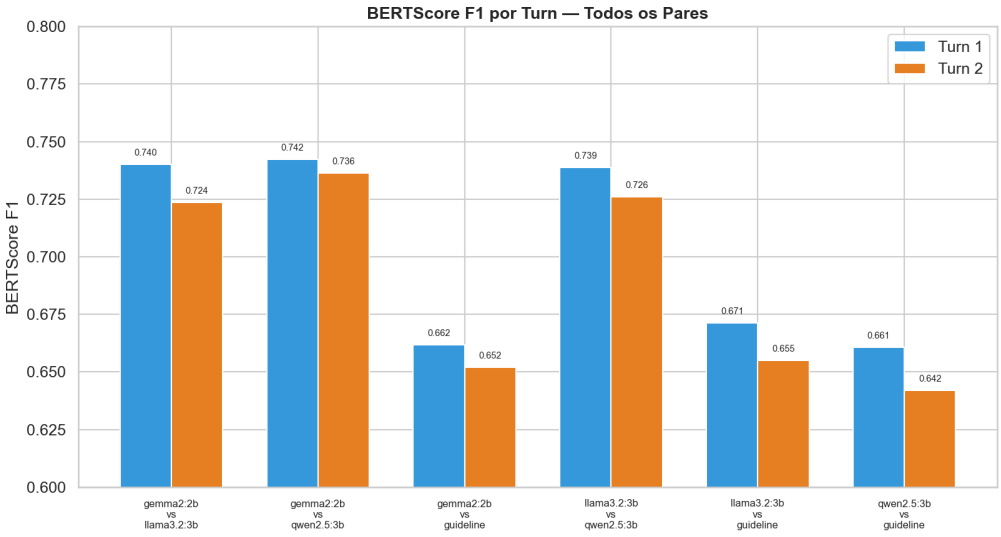

Figura 7: Distribuição do BERTScore F1 por turno para todos os pares

2.1.4 Resultados da avaliação com modelo juiz

Além das métricas automáticas de similaridade, um modelo juiz (GPT-4o-mini) foi utilizado para avaliar qualitativamente as respostas dissertativas. A Tabela 4 apresenta os scores médios atribuídos a cada modelo.

Tabela 4: Score médio atribuído pelo modelo juiz GPT-4o-mini.

Modelo	Score Médio
qwen2.5:3b	0,189
llama3.2:3b	0,139
gemma2:2b	0,120

A Figura 8 ilustra essa classificação. Os valores baixos de todos os modelos refletem a dificuldade geral de LLMs compactos diante de questões jurídicas dissertativas que exigem fundamentação técnica aprofundada.

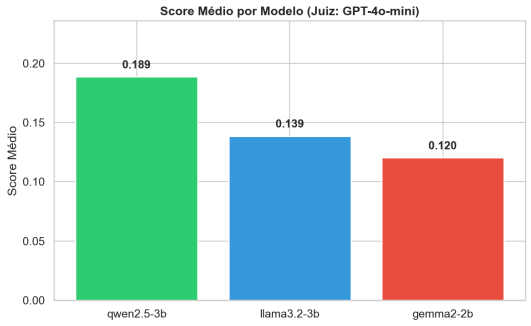

Figura 8: Score médio por modelo.

A Figura 9 ilustra que a desagregação por *turn* evidencia comportamentos distintos entre os modelos. O **llama3.2:3b** apresenta um aumento expressivo do Turn 1 (0,083) para o Turn 2 (0,205), enquanto o **qwen2.5:3b** exibe comportamento oposto, com redução no segundo *turn*. Esse resultado sugere diferenças na capacidade de aprofundamento das respostas entre os modelos.

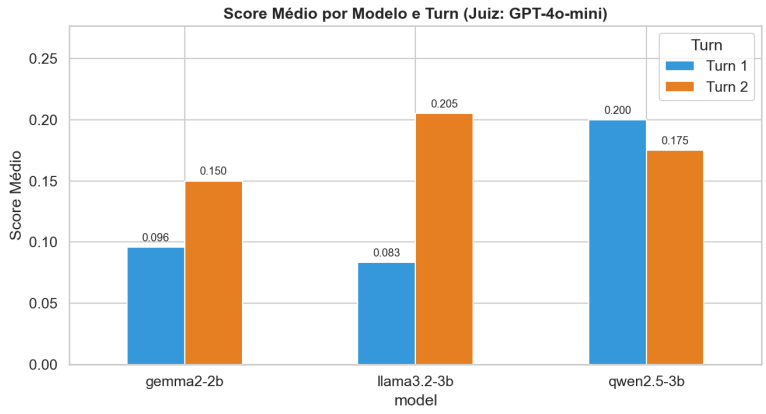

Figura 9: Score do juiz por modelo e turn.

O heatmap da Figura 10 detalha os scores por questão individual. Questões de direito administrativo obtiveram melhor desempenho geral, enquanto questões de direito do trabalho foram as mais difíceis para todos os modelos. Nota-se também que nenhum modelo domina todas as questões, evidenciando comportamentos complementares entre eles.

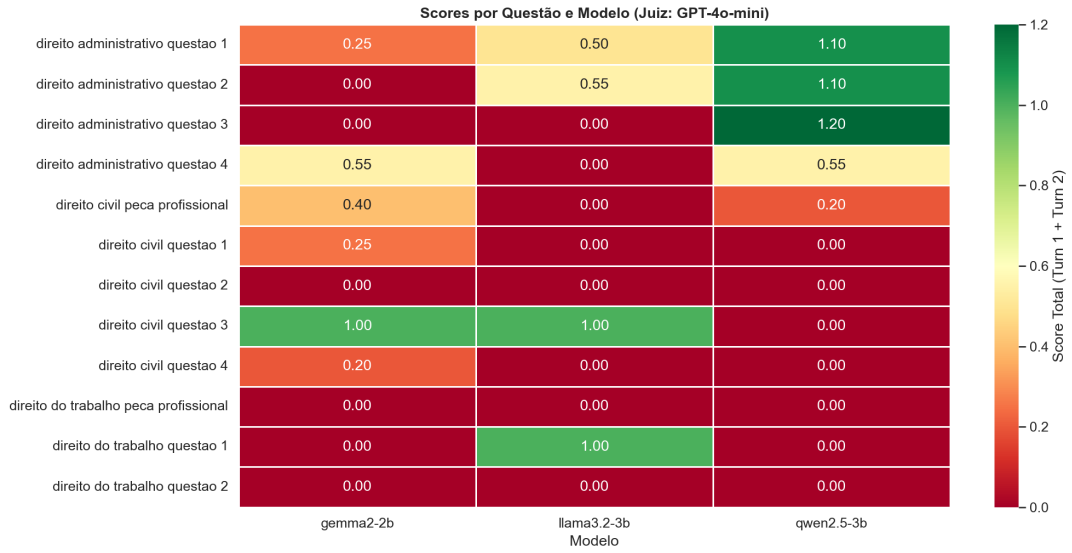

Figura 10: Scores do juiz por questão e modelo.

2.2 Fernanda Mirely Barbosa Souza

2.3 Éricles dos Santos Cunha

2.4 Júlia

2.5 Mikaela de Andrade Lima

2.6 Victor Leonardo Mascarenhas Soares Horta

3 Referências

1. Zhao, H. et al. *LLM Evaluation: A Comprehensive Survey.* arXiv, 2025. Disponível em: https://arxiv.org/html/2504.21202v1.
2. Maritaca AI. *OAB Bench.* Disponível em: https://github.com/maritaca-ai/oab-bench.
3. Garcia, E. *OAB Exams.* Disponível em: https://huggingface.co/datasets/eduagarcia/oab_exams.
4. Ollama. *Ollama Run LLMs locally.* Disponível em: https://ollama.com/.
5. Typer. *Typer library for building CLIs in Python.* Disponível em: https://typer.tiangolo.com/.
6. Astral. *uv Python package manager.* Disponível em: https://docs.astral.sh/uv/.