LinguagemSimples: Simplificação Automática de Decisões Judiciais com Modelos de Linguagem de Grande Escala

João Pedro Sansão, Michel Leles

https://doi.org/10.1590/SciELOPreprints.16575

Submetido em: 2026-06-16
Postado em: 2026-06-18 (versão 1)
(AAAA-MM-DD)

A moderação deste preprint recebeu o(s) endosso(s) de:

- Leonardo Araujo (ORCID: https://orcid.org/0000-0003-3884-2177)

LinguagemSimples: Simplificação Automática de Decisões Judiciais com Modelos de Linguagem de Grande Escala

João Pedro Hallack Sansão*[1], Michel Carlo Rodrigues Leles[2]

[1]Departamento de Tecnologia em Engenharia Civil, Computação, Automação, Telemática e Humanidades (DTECH), Universidade Federal de São João del-Rei, ORCID: https://orcid.org/0000-0003-0095-2629

[2]Departamento de Tecnologia em Engenharia Civil, Computação, Automação, Telemática e Humanidades (DTECH), Universidade Federal de São João del-Rei, ORCID: https://orcid.org/0000-0001-7399-7444

Resumo

A linguagem jurídica das decisões judiciais brasileiras — marcada por latinismos, jargões técnicos e orações subordinadas encadeadas — dificulta severamente a compreensão pelo cidadão comum. Este artigo apresenta o *LinguagemSimples*, um pipeline para simplificação automática de decisões judiciais utilizando modelos de linguagem de grande escala (LLMs). Foram avaliadas dezesseis técnicas — regras lexicais, Big Pickle (Few-Shot, Zero-Shot, CoT), Nemotron 3 Ultra (FS, ZS, CoT), DeepSeek V4 Flash (FS, ZS, CoT), Qwen 2.5 7B (FS, ZS, CoT), GPT-5.4 Mini (FS), GPT-5.4 (completo) (FS) e Gemini 3.5 Flash (FS) — sobre 100 decisões reais do STF nos temas consumidor, família e previdenciário. As métricas incluem legibilidade (Flesch Adaptado, Gunning-Fog), similaridade lexical (ROUGE) e preservação semântica (BERTScore). Adicionalmente, uma análise LLM-as-Judge (GPT-5.4 Mini) avaliou 1.500 saídas simplificadas em cinco categorias de erro. Todos os LLMs superam a baseline de regras, que reduziu a legibilidade ($-1,6$ pontos Flesch). DeepSeek V4 Flash e Big Pickle obtiveram os maiores ganhos de legibilidade (+24,3 pontos cada), enquanto o Qwen 2.5 7B Zero-Shot liderou em preservação semântica (BERTScore mBERT F1=0,748). O Chain-of-Thought mostrou-se contraproducente em todos os modelos, sendo o Few-Shot a estratégia de prompt mais eficaz. O GPT-5.4 Mini ofereceu o melhor custo-benefício entre latência e qualidade (+16,4 ganho Flesch, 0,697 BERTScore F1, ~2,5 s/doc), e o GPT-5.4 (completo) alcançou o maior ROUGE-1 (0,583) e o segundo maior BERTScore (0,713). A análise LLM-as-Judge revelou taxas de alucinação entre 7% (GPT-5.4 completo) e 49% (Qwen 2.5 7B FS), com perda de nuances como a categoria de erro mais frequente em todas as técnicas. O tema consumidor mostrou-se mais favorável à simplificação (+28,2 pontos), enquanto família foi o mais desafiador. O corpus e o código estão disponíveis publicamente.

Palavras-chave: Linguagem Simples, PLN Jurídico, Modelos de Linguagem, Decisões Judiciais, Avaliação de Simplificação

Abstract

The legal language of Brazilian court decisions — marked by Latinisms, technical jargon, and nested subordinate clauses — severely hinders comprehension by the average citizen. This paper presents *LinguagemSimples*, a pipeline for the automatic simplification of court decisions using large language models (LLMs). Sixteen techniques were evaluated — lexical rules, Big Pickle (Few-Shot, Zero-Shot, CoT), Nemotron 3 Ultra (FS, ZS, CoT), DeepSeek V4 Flash (FS, ZS, CoT), Qwen 2.5 7B (FS, ZS, CoT), GPT-5.4 Mini (FS), GPT-5.4 (full) (FS), and Gemini 3.5 Flash (FS) — on 100 real STF decisions across consumer, family, and social security law. Metrics include readability (Adapted Flesch, Gunning-Fog), lexical similarity (ROUGE), and semantic preservation (BERTScore).

*Autor correspondente

Additionally, an LLM-as-Judge analysis (GPT-5.4 Mini) evaluated 1,500 simplified outputs across five error categories. All LLMs outperform the rule-based baseline, which actually reduced readability (−1.6 Flesch points). DeepSeek V4 Flash and Big Pickle achieved the highest readability gains (+24.3 points each), while Qwen 2.5 7B Zero-Shot led in semantic preservation (BERTScore mBERT F1=0.748). Chain-of-Thought proved counterproductive across all models, with Few-Shot being the most effective prompting strategy. GPT-5.4 Mini offered the best latency-quality trade-off (+16.4 Flesch gain, 0.697 BERTScore F1, ~2.5 s/doc), and GPT-5.4 (full) achieved the highest ROUGE-1 (0.583) and second-highest BERTScore (0.713). The LLM-as-Judge analysis revealed hallucination rates ranging from 7% (GPT-5.4 full) to 49% (Qwen 2.5 7B FS), with nuance loss being the most frequent error category across all techniques. Consumer law proved the most favorable domain for simplification (+28.2 points), while family law was the most challenging. The corpus and code are publicly available.

Keywords: Plain Language, Legal NLP, Language Models, Court Decisions, Simplification Evaluation

1 Introdução

O direito de compreender a própria decisão judicial é um desdobramento fundamental do Estado Democrático de Direito. A Constituição Federal de 1988, em seu artigo 93, inciso IX, exige a fundamentação de todas as decisões judiciais [1]. No entanto, a linguagem empregada nessas decisões — marcada por latinismos (*data venia*, *in casu*, *ex vi*), jargões técnicos e construções sintáticas com múltiplas orações subordinadas — torna o texto inacessível ao jurisdicionado não especialista.

O movimento da Linguagem Simples (*Plain Language*) ganhou força no Brasil nos últimos anos. O Conselho Nacional de Justiça (CNJ) instituiu o Pacto do Judiciário pela Linguagem Simples (Portaria CNJ n. 351/2023) [2], incentivando magistrados a adotar comunicação mais clara e acessível. Diversos tribunais têm implementado programas de simplificação, como o TJ Mais Simples do Tribunal de Justiça de São Paulo e o Laboratório de Linguagem Simples do Tribunal Regional Federal da 1ª Região. Paralelamente, o relatório Justiça em Números 2022 [3] aponta que o Judiciário brasileiro recebeu mais de 27 milhões de novos casos naquele ano, dos quais grande parte envolve cidadãos sem representação advocatícia regular — reforçando a urgência de tornar a linguagem judicial acessível.

O desafio de compreensão é agravado pelo perfil educacional da população brasileira. Segundo o Indicador de Alfabetismo Funcional (INAF) 2022 [4], apenas 12% da população brasileira pode ser considerada plenamente alfabetizada, enquanto 29% apresenta nível elementar de alfabetismo — capazes de ler títulos ou frases curtas, mas com grande dificuldade diante de textos longos com vocabulário especializado. A PNAD Contínua 2021 [5] revela que 77% dos brasileiros com 25 anos ou mais não possuem ensino superior completo. Esses dados evidenciam que a linguagem rebuscada das decisões judiciais não é meramente um desconforto estético, mas uma barreira real ao exercício da cidadania.

Os avanços em Processamento de Linguagem Natural (PLN), especialmente com os Modelos de Linguagem de Grande Escala (LLMs) [6, 7], abriram novas possibilidades para simplificação textual automatizada [8, 9]. Modelos como GPT-4 [10], Llama 2 [11], Mixtral [12] e outros demonstraram capacidade de reescrever textos complexos em linguagem mais acessível, mantendo o conteúdo informacional original [13]. A viabilidade técnica desses modelos para o domínio jurídico tem sido investigada em contextos internacionais [14, 15], mas estudos focados no português brasileiro ainda são escassos [16].

Este trabalho apresenta o *LinguagemSimples*, um pipeline computacional para simplificação automática de decisões judiciais brasileiras. As principais contribuições são:

1. Um corpus de 100 decisões reais do STF nos temas consumidor, família e previdenciário, pareadas com versões simplificadas por dezesseis técnicas distintas;
2. Uma comparação sistemática entre técnica baseada em regras lexicais e quinze baseadas em LLMs (Big Pickle com três modos, Nemotron 3 Ultra com três modos, DeepSeek V4 Flash com três modos, Qwen 2.5 7B com três modos, GPT-5.4 Mini, GPT-5.4 e Gemini 3.5 Flash);
3. Uma avaliação multidimensional com métricas de legibilidade, similaridade textual e preservação semântica;

4. A análise de diferentes estratégias de engenharia de prompt (zero-shot, few-shot, chain-of-thought) para o domínio jurídico;
5. Disponibilização pública do corpus e do código-fonte para reproducibilidade.

Ressalta-se que o objetivo não é identificar permanentemente o ‘‘melhor LLM’’ — os modelos aqui testados refletem o estado da arte em meados de 2026 e serão inevitavelmente sucedidos — mas compreender como diferentes estratégias de simplificação (regras, few-shot, zero-shot, chain-of-thought) afetam o equilíbrio entre legibilidade e preservação semântica, gerando conhecimento transferível para futuras arquiteturas e modelos.

O restante deste artigo está organizado como segue. A Seção 2 discute trabalhos relacionados. A Seção 3 descreve o corpus, as técnicas de simplificação e as métricas de avaliação. A Seção 4 apresenta os resultados quantitativos e qualitativos. A Seção 5 discute os achados e limitações. A Seção 6 conclui o artigo com direções para trabalhos futuros.

2 Trabalhos Relacionados

A simplificação de textos jurídicos situa-se na interseção de duas áreas de pesquisa: a Linguística Computacional aplicada ao domínio jurídico (Legal NLP) e o movimento da Linguagem Simples (Plain Language Movement).

2.1 Plain Language Movement no Brasil e no Mundo

O movimento internacional da Linguagem Simples remonta à década de 1970, com a Plain English Campaign no Reino Unido e o Plain Writing Act nos Estados Unidos [17]. No Brasil, o movimento ganhou institucionalidade recente com o Pacto do Judiciário pela Linguagem Simples do CNJ [2], que recomenda o uso de frases curtas, ordem direta e vocabulário acessível nas decisões judiciais. O direito fundamental à compreensão das decisões judiciais encontra fundamento nos princípios constitucionais da publicidade e do acesso à justiça. Análises linguísticas de acórdãos de tribunais superiores brasileiros revelam que a maioria das sentenças ultrapassa 30 palavras por período e emprega vocabulário jurídico especializado de forma generalizada. Estudos empíricos indicam que cidadãos com baixa escolaridade compreendem menos de um terço do conteúdo de decisões judiciais simples, evidenciando o impacto da linguagem processual na efetivação do acesso à justiça.

Internacionalmente, o movimento Plain Language possui marcos legais mais consolidados. O Plain Writing Act de 2010 nos Estados Unidos determina que agências federais utilizem linguagem clara em todos os documentos voltados ao público. No Reino Unido, a Plain English Campaign atua desde 1979, certificando documentos governamentais e empresariais. No Brasil, tais iniciativas são mais recentes e ainda carecem de ferramentas tecnológicas que auxiliem na aplicação em larga escala, lacuna que este trabalho busca contribuir para preencher.

2.2 Simplificação Textual com PLN

A simplificação textual automática tem sido estudada sob diferentes paradigmas. Niklaus et al. [8] apresentam uma revisão abrangente, categorizando as abordagens em baseadas em regras, baseadas em corpus e neurais. Shardlow [9] estabelecem as bases metodológicas para avaliação, distinguindo entre simplificação lexical, sintática e semântica.

No contexto jurídico internacional, [18] investigam o uso de linguagem simples em contratos jurídicos, destacando o potencial de ferramentas automáticas para aumentar a legibilidade sem comprometer a precisão terminológica. Chalkidis et al. [15] apresentam o modelo LEGAL-BERT, destacando a escassez de recursos de PLN jurídico para línguas além do inglês.

No Brasil, Silveira et al. [16] propuseram o LegalBert-pt, um modelo de linguagem pré-treinado especificamente para processamento de textos jurídicos brasileiros. Araujo et al. [19] apresentaram o LeNER-Br, um conjunto de dados e modelo para reconhecimento de entidades nomeadas no domínio

jurídico brasileiro. Manor and Li [18] exploram o uso de linguagem simples em documentos jurídicos, analisando o impacto na compreensibilidade de contratos.

Recentemente, com a popularização dos LLMs, novos estudos têm investigado seu potencial para simplificação jurídica. [14] investigaram o uso de GPT-4 para anotação semântica de textos jurídicos, demonstrando sua eficácia em tarefas do domínio legal. Paula and Camilo-Junior [20] avaliam a simplificação de decisões judiciais brasileiras com LLMs, analisando o equilíbrio entre legibilidade e preservação semântica. [21] discutem o desempenho de modelos GPT em tarefas jurídicas, abrindo caminho para aplicações de simplificação textual no âmbito forense. Diferentes técnicas de engenharia de prompt, como zero-shot, few-shot, CoT e prompting iterativo, têm sido exploradas para tarefas de PLN jurídico. A sumarização automática de acórdãos judiciais tem sido investigada utilizando métodos extrativos baseados em frequência de termos e posição de sentenças. O projeto PorSimples [22] propôs um corpus paralelo e ferramentas para simplificação textual em português brasileiro, abrangendo textos jornalísticos e literários. O modelo BERTimbau [23] foi desenvolvido especificamente para o português brasileiro, demonstrando a viabilidade de modelos de linguagem específicos para o idioma.

2.3 Métricas de Avaliação

Em relação a métricas de avaliação, Lin [24] propôs o ROUGE (Recall-Oriented Understudy for Gisting Evaluation), baseado em sobreposição de n-gramas, amplamente utilizado para sumarização e simplificação. Zhang et al. [25] introduziram o BERTScore, que utiliza embeddings contextuais para medir similaridade semântica, mostrando correlação mais forte com avaliação humana do que métricas baseadas em n-gramas. Para o português brasileiro, [26] estabeleceu índices de legibilidade para o português, adaptando métricas como o índice Flesch para o idioma. O BERTScore tem sido validado para o português brasileiro, demonstrando sua eficácia em tarefas de geração de texto. O modelo BERTimbau [23] foi avaliado comparativamente a outros modelos BERT para o português brasileiro — como `bert-base-multilingual-cased` — fornecendo subsídios para a escolha do modelo de embeddings.

Este trabalho se diferencia dos anteriores por (i) focar exclusivamente em decisões judiciais reais brasileiras do STF, (ii) comparar múltiplos LLMs disponíveis gratuitamente via API unificada (OpenCode Zen), (iii) avaliar três estratégias de engenharia de prompt (zero-shot, few-shot, chain-of-thought) em quatro modelos (Big Pickle, Nemotron 3 Ultra, DeepSeek V4 Flash e Qwen 2.5 7B), (iv) utilizar métricas complementares de legibilidade (Flesch Adaptado, Gunning-Fog), similaridade lexical (ROUGE-1/2/L) e preservação semântica (BERTScore) para uma avaliação multidimensional, e (v) disponibilizar publicamente o corpus pareado e o código-fonte para reproducibilidade.

3 Metodologia

A metodologia compreende três etapas principais: construção do corpus, aplicação das técnicas de simplificação e avaliação multidimensional. A Figura 1 apresenta uma visão geral do pipeline.

3.1 Corpus de Decisões Judiciais

O corpus foi construído a partir do dataset público `celsowm/jurisprudencias_stf` disponível no HuggingFace Datasets [27]. Este dataset contém 29.398 decisões do Supremo Tribunal Federal (STF) com metadados como número do processo, ementa, tese, relator, data de publicação e classe processual.

Foram selecionadas 100 decisões — 34 de consumidor, 33 de família e 33 de previdenciário, balanceadas entre os três temas jurídicos de maior relevância social: consumidor, família e previdenciário. A classificação temática foi realizada automaticamente por meio de um dicionário de palavras-chave específicas para cada tema. Por exemplo, para o tema consumidor, as palavras-chave incluíram: *consumidor*, *plano de saúde*, *fornecedor*, *relação de consumo* e *Código de Defesa do Consumidor*. Para família: *divórcio*, *guarda*, *pensão alimentícia*, *casamento* e *união estável*. Para previdenciário: *aposentadoria*, *INSS*, *auxílio-doença*, *pensão por morte* e *benefício previdenciário*.

Figura 1: Visão geral do pipeline do LinguagemSimples.

Após a classificação temática, os documentos foram ordenados por data de publicação (mais recentes primeiro) e selecionaram-se os primeiros N por tema até atingir o balanceamento desejado (34 consumidor, 33 família, 33 previdenciário). Uma verificação manual rápida eliminou documentos duplicados ou mal classificados (e.g., ementas que mencionavam ‘‘consumidor’’ em contexto não jurídico).

Cada documento foi submetido a um processo de limpeza textual mínimo: remoção de espaços duplicados e normalização de caracteres Unicode. Não foram removidos termos jurídicos *a priori*, para que a simplificação pudesse ser avaliada sobre o texto integral da decisão.

O corpus final apresenta as seguintes estatísticas: o texto original médio possui 527 palavras, variando de 46 a 3.030 palavras. A sentença média no original contém 28,5 palavras. O índice Flesch Adaptado médio dos originais é 39,7 (classificado como ‘‘muito difícil’’), confirmando a complexidade linguística das decisões selecionadas.

A Tabela 1 apresenta a composição detalhada do corpus.

Tabela 1: Composição do corpus de decisões do STF

Tema	Documentos	Média caracteres	Fonte
Consumidor	34	2975	STF (HuggingFace)
Família	33	5242	STF (HuggingFace)
Previdenciário	33	2348	STF (HuggingFace)
Total	**100**	**3516**	–

3.2 Técnicas de Simplificação

Dezesseis técnicas de simplificação foram implementadas e avaliadas.

3.2.1 Baseado em Regras

A técnica baseada em regras (‘‘Regras’’) implementa três operações sequenciais, servindo como baseline mínimo:

1. **Substituição lexical**: dicionário de 30 termos jurídicos mapeados para equivalentes em linguagem simples. Exemplos: *in casu* → *neste caso*; *data venia* → *com respeito*; *decisum* → *decisão*; *ex vi* → *nos termos de*; *quo* → *em questão*.

Tabela 2: Resultados comparativos das técnicas de simplificação (Flesch original médio = 39,7; Gunning-Fog original médio = 21.7)

Modelo	Modo	Flesch Simpl.	ΔFlesch	Fog	ROUGE			BS mBERT	BS BERTimbau
					R-1	R-2	R-L		
	Few-Shot	53.6	+13.9	18.3	0.4614	0.2603	0.3513	0.7259	0.6731
Qwen 2.5 7B	Zero-Shot	53.3	+13.6	18.3	0.4888	0.2814	0.3726	0.7476	0.6986
	CoT	47.0	+7.2	19.0	0.4156	0.2228	0.2516	0.6772	0.6140
	Few-Shot	64.0	+24.3	18.0	0.3936	0.1485	0.2301	0.6718	0.5858
Big Pickle	Zero-Shot	63.4	+23.7	17.9	0.4332	0.1694	0.2418	0.6807	0.5912
	CoT	62.0	+22.2	18.4	0.3361	0.1157	0.1870	0.6551	0.5595
	Few-Shot	62.4	+22.7	17.4	0.4757	0.2336	0.3209	0.7038	0.6281
Nemotron 3 Ultra	Zero-Shot	63.0	+23.3	17.5	0.4667	0.2046	0.2645	0.6835	0.5968
	CoT	60.5	+20.7	18.0	0.3843	0.1481	0.2159	0.6635	0.5703
	Few-Shot	64.1	+24.3	18.1	0.3961	0.1490	0.2300	0.6729	0.5862
DeepSeek V4 Flash	Zero-Shot	61.6	+21.9	18.3	0.4419	0.1760	0.2485	0.6818	0.5970
	CoT	61.6	+21.8	18.6	0.3332	0.1109	0.1881	0.6518	0.5590
GPT-5.4 Mini	Few-Shot	56.1	+16.4	19.3	0.4584	0.2400	0.3222	0.6969	0.6381
GPT-5.4 (completo)	Few-Shot	57.3	+17.6	18.5	0.5828	0.3274	0.4108	0.7127	0.6416
Gemini 3.5 Flash	Few-Shot	61.5	+21.7	19.2	0.2547	0.0943	0.1599	0.6459	0.5419
Baseado em Regras	—	38.1	-1.6	21.9	0.9919	0.9767	0.9919	0.9940	0.9940

Tabela 3: Ganho Flesch por tema e técnica (Flesch original médio por tema)

Modelo	Modo	Consumidor	Família	Previdenciário
Original	—	36.9	41.7	40.8
	Few-Shot	+16.7	+8.4	+16.5
Qwen 2.5 7B	Zero-Shot	+15.9	+12.1	+12.7
	CoT	+10.1	+7.5	+4.0
	Few-Shot	+27.4	+22.9	+22.5
Big Pickle	Zero-Shot	+26.7	+22.8	+21.4
	CoT	+26.9	+19.2	+20.5
	Few-Shot	+25.5	+21.6	+20.9
Nemotron 3 Ultra	Zero-Shot	+27.5	+21.3	+21.0
	CoT	+23.0	+19.6	+19.5
	Few-Shot	+28.2	+21.7	+23.0
DeepSeek V4 Flash	Zero-Shot	+24.4	+20.3	+20.8
	CoT	+26.1	+19.0	+20.2
GPT-5.4 Mini	Few-Shot	+19.5	+17.1	+12.4
GPT-5.4 (completo)	Few-Shot	+19.9	+16.3	+16.4
Gemini 3.5 Flash	Few-Shot	+25.0	+22.1	+18.0
Regras	—	-1.8	-1.8	-1.3

Tabela 4: BERTScore por técnica e modelo de avaliação

Modelo	Modo	mBERT (multilingual)			BERTimbau (pt-br)		
		P	R	F1	P	R	F1
Qwen 2.5 7B	Few-Shot	0.7436	0.7100	0.7259	0.7030	0.6486	0.6731
	Zero-Shot	0.7623	0.7342	0.7476	0.7246	0.6762	0.6986
	CoT	0.6874	0.6681	0.6772	0.6301	0.6009	0.6140
Big Pickle	Few-Shot	0.6888	0.6563	0.6718	0.6037	0.5704	0.5858
	Zero-Shot	0.6892	0.6731	0.6807	0.5956	0.5883	0.5912
	CoT	0.6764	0.6361	0.6551	0.5840	0.5394	0.5595
Nemotron 3 Ultra	Few-Shot	0.7131	0.6957	0.7038	0.6345	0.6240	0.6281
	Zero-Shot	0.6867	0.6811	0.6835	0.5935	0.6019	0.5968
	CoT	0.6729	0.6550	0.6635	0.5768	0.5655	0.5703
DeepSeek V4 Flash	Few-Shot	0.6912	0.6564	0.6729	0.6045	0.5708	0.5862
	Zero-Shot	0.6908	0.6737	0.6818	0.6012	0.5943	0.5970
	CoT	0.6759	0.6302	0.6518	0.5887	0.5338	0.5590
GPT-5.4 Mini	Few-Shot	0.7210	0.6753	0.6969	0.6719	0.6094	0.6381
GPT-5.4 (completo)	Few-Shot	0.7181	0.7085	0.7127	0.6406	0.6450	0.6416
Gemini 3.5 Flash	Few-Shot	0.6833	0.6134	0.6459	0.5904	0.5037	0.5419

Tabela 5: Índice Gunning-Fog por tema e técnica (Fog e variação percentual)

Modelo	Modo	Consumidor		Família		Previdenciário	
		Fog	Variação (%)	Fog	Variação (%)	Fog	Variação (%)
Fog. Orig.	—	22.2	—	21.9	—	21.1	—
Qwen 2.5 7B	Few-Shot	18.6	-16.1%	18.9	-13.4%	17.2	-18.3%
	Zero-Shot	18.7	-16.1%	18.5	-15.4%	17.7	-16.2%
	CoT	19.1	-13.9%	19.0	-13.1%	19.0	-10.1%
Baseado em Regras	—	22.4	+0.9%	22.1	+0.8%	21.2	+0.7%
Big Pickle	Few-Shot	18.1	-18.5%	17.9	-18.2%	17.9	-15.4%
	Zero-Shot	17.9	-19.2%	17.8	-18.9%	17.9	-15.3%
	CoT	18.3	-17.7%	18.8	-14.3%	18.3	-13.4%
Nemotron 3 Ultra	Few-Shot	17.7	-20.5%	17.3	-20.8%	17.2	-18.4%
	Zero-Shot	17.4	-21.9%	17.6	-19.4%	17.6	-16.8%
	CoT	18.4	-17.1%	17.7	-18.9%	17.7	-16.0%
DeepSeek V4 Flash	Few-Shot	18.2	-18.3%	18.5	-15.6%	17.8	-15.6%
	Zero-Shot	18.4	-17.2%	18.3	-16.3%	18.1	-14.3%
	CoT	18.6	-16.5%	18.9	-13.9%	18.5	-12.2%
GPT-5.4 Mini	Few-Shot	19.4	-12.5%	19.0	-13.2%	19.5	-7.6%
GPT-5.4 (completo)	Few-Shot	18.8	-15.5%	18.6	-15.1%	18.1	-14.0%
Gemini 3.5 Flash	Few-Shot	19.3	-13.3%	18.9	-13.9%	19.4	-8.1%

Tabela 6: Exemplos qualitativos — Consumidor (RE 1499539 RG)

Documento	**Texto**
Original	Ementa: Direito tributário. Diferencial de Alíquota de ICMS. Operações Interestaduais. Consumidor final contribuinte. Lei complementar. Matéria infraconstitucional. I. Caso em exame 1. Recurso extraordinário de acórdão do Tribunal de Justiça do Estad...
Qwen 2.5 7B (FS)	Ementa: Imposto. Diferencial de alíquota. Compras entre estados. Consumidor que paga imposto. Lei complementar. Questão menor. I. Caso em exame 1. Recurso de um tribunal que disse que é preciso cobrar diferencial de alíquota de imposto (ICMS-DIFAL) e...
Qwen 2.5 7B (ZS)	Texto simplificado: Ementa: Direito tributário. Diferencial de alíquota do ICMS. Vendas entre estados. Cliente final que paga o imposto. Lei complementar. Assunto não constitucional. 1. Caso em exame: - Recurso do Tribunal de Justiça de Minas Gerai...
Qwen 2.5 7B (CoT)	1. Identificação das partes envolvidas: - Não há partes específicas identificadas como ”pedindo”ou ”acusadas”no texto original. Trata-se de uma discussão sobre a aplicabilidade de uma lei. 2. Decisão principal: - A decisão principal é que a ...
Big Pickle (FS)	Aqui está a explicação da decisão do STF sobre o ICMS-DIFAL em linguagem simples, seguindo as instruções: **Decisão do STF sobre o ICMS-DIFAL explicada** **O que é o ICMS-DIFAL?** É a **diferença de alíquota** do ICMS. Esse imposto pode ser co...
Nemotron	**Assunto:** Direito tributário. Diferença de alíquota do ICMS (ICMS-DIFAL). Compras entre estados. Consumidor final que paga ICMS. Lei Complementar nº 87/1996. Questão que não é constitucional. **O caso:** Um recurso foi enviado ao Supremo Tribunal...
GPT-5.4 Mini	Este caso trata de direito tributário, ou seja, de impostos. O Tribunal de Justiça de Minas Gerais entendeu que o ICMS-DIFAL pode ser cobrado em vendas feitas para outro estado, quando o comprador final também é contribuinte do ICMS. Isso porque a L...
GPT-5.4 (completo)	**Texto simplificado:** **Resumo da decisão** Este caso é da área de **direito tributário** e trata da cobrança do **diferençal de alíquota do ICMS (ICMS-DIFAL)** em vendas feitas entre estados. **1. O que aconteceu no caso** Foi apresentado um **...
Gemini 3.5 Flash	**O que aconteceu:** A Justiça de Minas Gerais decidiu que o estado pode cobrar uma diferença de imposto (chamada de ICMS-DIFAL) quando uma empresa compra mercadorias de outro estado para consumo próprio. A decisão se baseou na Lei Complementar nº 87...
Regras	Ementa: Direito tributário. Diferencial de Alíquota de ICMS. Operações Interestaduais. Consumidor final contribuinte. Lei complementar. Matéria infraconstitucional. I. Caso em exame 1. Recurso extraordinário de acórdão do Tribunal de Justiça do Estad...

Tabela 7: Exemplos qualitativos — Família (RE 842844)

Documento	Texto
Original	EMENTA: RECURSO EXTRAORDINÁRIO. REPERCUSSÃO GERAL RECONHECIDA. CONSTITUCIONAL. ADMINISTRATIVO. SERVIDORA GESTANTE. GRAVIDEZ DURANTE O PERÍODO DA PRESTAÇÃO DOS SERVIÇOS. VÍNCULO COM A ADMINISTRAÇÃO PÚBLICA POR CARGO COMISSIONADO, NÃO EFETIVO, OU POR C...
Qwen 2.5 7B (FS)	**RECURSO EXTRAORDINÁRIO. REPERCUSSÃO GERAL RECONHECIDA.** **CONSTITUCIONAL. ADMINISTRATIVO.** **Servidora Gestante.** **Licença Maternidade e Estabilidade Provisória.** **Artigos 7º, XVIII, e 10, II, B, da Constituição Federal.** **Reafirmação ...
Qwen 2.5 7B (ZS)	**EMENTA: RECURSO EXTRAORDINÁRIO. REPERCUSSÃO GERAL RECONHECIDA. CONSTITUCIONAL. ADMINISTRATIVO. SERVIDORA GESTANTE. GRAVIDEZ DURANTE O PERÍODO DA PRESTAÇÃO DOS SERVIÇOS. VÍNCULO COM A ADMINISTRAÇÃO PÚBLICA POR CARGO COMISSIONADO, NÃO EFETIVO, OU POR...
Qwen 2.5 7B (CoT)	1. **Partes envolvidas:** - **Trabalhadora gestante:** A pessoa que está grávida e trabalha para o governo. - **Administração Pública:** O governo que emprega a trabalhadora gestante. 2. **Decisão principal:** - O Supremo Tribunal Federal d...
Big Pickle (FS)	Texto simplificado: O Supremo Tribunal Federal (STF) decidiu que **toda servidora pública grávida tem direito à licença-maternidade e à estabilidade provisória**, mesmo que ela não seja concursada. Isso vale para quem ocupa cargo de confiança (comis...
Nemotron	**Resumo da Decisão do Supremo Tribunal Federal (STF)** **O que foi decidido:** Toda trabalhadora grávida que presta serviço para o governo tem direito à **licença-maternidade** e à **estabilidade no emprego** (não pode ser demitida sem justa causa)...
GPT-5.4 Mini	Neste caso, a Justiça analisou o recurso de um processo sobre uma servidora grávida que trabalhava para a Administração Pública. A gravidez aconteceu enquanto ela ainda prestava serviços ao órgão público. O ponto principal era saber se ela tinha dir...
GPT-5.4 (completo)	**Texto simplificado:** Este caso trata de uma servidora pública grávida. Ela trabalhou para a Administração Pública em **cargo comissionado**, **cargo não efetivo** ou por **contrato temporário**. O Supremo Tribunal Federal decidiu que ela tem dire...
Gemini 3.5 Flash	**Decisão do Supremo Tribunal Federal (STF):** O STF decidiu que toda trabalhadora grávida que presta serviços para a administração pública tem direito à licença-maternidade e à estabilidade provisória (garantia de não ser demitida sem justa causa d...
Regras	EMENTA: RECURSO EXTRAORDINÁRIO. REPERCUSSÃO GERAL RECONHECIDA. CONSTITUCIONAL. ADMINISTRATIVO. SERVIDORA GESTANTE. GRAVIDEZ DURANTE O PERÍODO DA PRESTAÇÃO DOS SERVIÇOS. VÍNCULO COM A ADMINISTRAÇÃO PÚBLICA POR CARGO COMISSIONADO, NÃO EFETIVO, OU POR C...

Tabela 8: Exemplos qualitativos — Previdenciário (RE 626837)

Documento	**Texto**
Original	EMENTA Recurso extraordinário. Repercussão geral. Tributário. Contribuição previdenciária. Imunidade recíproca. Inexistência. Artigo 195, I, a, e II, da CF, na versão da EC nº 20/98. Lei nº 10.887/04. Exercentes de mandato eletivo. Agentes políticos....
Qwen 2.5 7B (FS)	**Texto simplificado:** 1. A imunidade recíproca do artigo 150, VI, a, da Constituição Federal só se aplica a impostos. Na ADI nº 2.024/DF, a Corte decidiu que essa imunidade não pode ser usada para evitar a contribuição previdenciária. 2. Na decis...
Qwen 2.5 7B (ZS)	**Texto simplificado:** Este recurso foi analisado e negado. O caso envolve a questão de se os exercentes de mandatos eletivos (como deputados e senadores) precisam pagar uma contribuição previdenciária sobre seus salários recebidos pelo governo. 1...
Qwen 2.5 7B (CoT)	1. **Identificação das partes envolvidas:** - **Parte que pede:** Os exercentes de mandato eletivo (como prefeitos, deputados, senadores, etc.). - **Parte requerida:** O governo federal, estados, Distrito Federal e municípios. 2. **Decisão pri...
Big Pickle (FS)	Neste caso, o Supremo Tribunal Federal (STF) decidiu se os políticos eleitos (vereadores, prefeitos, governadores e deputados) precisam pagar contribuição ao INSS sobre o salário que recebem do governo. O tribunal entendeu que **sim, eles precisam p...
Nemotron	**Texto simplificado:** **Resumo do julgamento:** Recurso extraordinário com repercussão geral. Tema: contribuição previdenciária e imunidade recíproca. **O que foi decidido:** Não existe imunidade recíproca para a contribuição previdenciária. A im...
GPT-5.4 Mini	**Texto simplificado:** O caso trata de uma discussão sobre contribuição previdenciária, isto é, o valor pago para a Previdência Social. O Supremo Tribunal Federal decidiu que a regra de imunidade entre entes públicos não vale para contribuição pre...
GPT-5.4 (completo)	**Texto simplificado:** Este caso trata da cobrança de **contribuição previdenciária** sobre os valores pagos a pessoas que exercem **mandato eletivo**, como prefeitos, vereadores, governadores e outros agentes políticos. O Supremo Tribunal Federal...
Gemini 3.5 Flash	**Resumo do caso:** O Supremo Tribunal Federal (STF) decidiu que políticos eleitos (como prefeitos, deputados e vereadores) devem pagar a contribuição do INSS sobre os salários que recebem do governo. Essa decisão tem "repercussão geral", o que signi...
Regras	EMENTA Recurso extraordinário. Repercussão geral. Tributário. Contribuição previdenciária. Imunidade recíproca. Inexistência. Artigo 195, I, a, e II, da CF, na versão da EC nº 20/98. Lei nº 10.887/04. Exercentes de mandato eletivo. Agentes políticos....

2. **Divisão de sentenças**: sentenças com mais de 40 palavras são divididas em períodos menores, utilizando conectivos (*que*, *qual*, *onde*) e pontuação como pontos de corte.
3. **Voz ativa**: conversão de construções na voz passiva para voz ativa, quando o agente da ação é identificável.

3.2.2 Modelos de Linguagem (LLMs)

Sete famílias de LLMs foram acessadas via API, todos com temperatura 0,3 e máximo de 8192 tokens de saída. A temperatura 0,3 foi escolhida como compromisso entre determinismo e diversidade lexical: estudos piloto com 5 documentos mostraram que temperatura 0,0 produzia saídas excessivamente repetitivas (mesmas frases de abertura em todos os documentos), enquanto temperaturas acima de 0,5 introduziam variação excessiva entre execuções; o valor 0,3 oferece leve variação lexical sem comprometer a comparabilidade entre técnicas. Três deles (Big Pickle, Nemotron 3 Ultra, DeepSeek V4 Flash) compartilham a mesma rota de API do OpenCode Zen[1], que oferece interface compatível com OpenAI, simplificando a integração. O Qwen 2.5 7B, o GPT-5.4 Mini, o GPT-5.4 (completo) e o Gemini 3.5 Flash foram acessados via OpenRouter[2]. As chamadas via OpenCode Zen foram gratuitas; as chamadas via OpenRouter utilizaram o serviço regular pago dos provedores.

- **Big Pickle**: modelo de raciocínio (*reasoning model*) proprietário do OpenCode, disponível gratuitamente desde outubro de 2025[3]. Possui arquitetura otimizada para tarefas que exigem planejamento e execução multi-etapas, com janela de contexto de 200 mil tokens. Utiliza processamento interno de cadeia de pensamento antes de gerar a resposta final, o que tende a produzir textos mais elaborados e semanticamente fiéis ao original.
- **Nemotron 3 Ultra Free**: modelo da família Nemotron (NVIDIA), lançado em junho de 2026[4]. Possui arquitetura híbrida *Mamba-Transformer* com Mistura de Especialistas (*Mixture-of-Experts*, MoE), totalizando 550 bilhões de parâmetros, dos quais aproximadamente 55 bilhões são ativados por token. Suporta até 1 milhão de tokens de contexto e foi projetado para *reasoning* e orquestração de agentes de longa duração. É multilíngue e possui pesos abertos sob licença NVIDIA Open Model License.
- **DeepSeek V4 Flash Free**: modelo da família DeepSeek (China), lançado em abril de 2026[5]. Também utiliza arquitetura MoE, com 284 bilhões de parâmetros totais e 13 bilhões ativados por token, oferecendo alta eficiência computacional. Suporta 1 milhão de tokens de contexto e três modos de raciocínio (rápido, equilibrado e profundo). É distribuído sob licença MIT e otimizado para tarefas de alta frequência e baixa latência.
- **GPT-5.4 Mini**: modelo da OpenAI otimizado para tarefas de alta vazão, lançado em março de 2026[6]. Possui janela de contexto de 400 mil tokens e é acessado via OpenRouter com API compatível com OpenAI. Oferece equilíbrio entre capacidade e custo, com preço de US$ 0,75/milhão de tokens de entrada e US$ 4,50/milhão de tokens de saída. Suporta chamada de funções, raciocínio multi-etapas e saída estruturada.
- **GPT-5.4 (completo)**: modelo *frontier* completo da OpenAI, acessado via OpenRouter[7]. Unifica as linhas Codex e GPT em um único sistema, com janela de contexto de 1 milhão de tokens (922 mil de entrada, 128 mil de saída) e suporte a texto e imagem. Oferece preço de US$ 2,50/milhão de tokens de entrada e US$ 15/milhão de tokens de saída. Em comparação com o GPT-5.4 Mini — que possui janela de 400 mil tokens, preço de US$ 0,75/milhão de entrada e US$ 4,50/milhão de saída e é otimizado para alta vazão — o modelo completo oferece maior capacidade de raciocínio e contexto mais longo, com custo aproximadamente 3,3 vezes maior.

[1] https://opencode.ai/zen
[2] https://openrouter.ai
[3] https://opencode.ai/zen
[4] https://research.nvidia.com/labs/nemotron/Nemotron-3-Ultra
[5] https://api-docs.deepseek.com
[6] https://openrouter.ai/openai/gpt-5.4-mini
[7] https://openrouter.ai/openai/gpt-5.4

- **Gemini 3.5 Flash**: modelo da Google acessado via OpenRouter[8]. Otimizado para tarefas de alta vazão com baixa latência, oferece preço de US$ 1,50/milhão de tokens de entrada e US$ 9,00/milhão de tokens de saída, com janela de contexto de 1 milhão de tokens.
- **Qwen 2.5 7B Instruct**: modelo de código aberto da Alibaba Cloud, acessado via OpenRouter[9]. Com apenas 7 bilhões de parâmetros, representa uma alternativa leve e de baixo custo (US$ 0,07/milhão de tokens de entrada e US$ 0,21/milhão de tokens de saída). Sua inclusão permite avaliar o desempenho de modelos menores e mais acessíveis em relação aos LLMs de grande porte. Assim como Big Pickle, Nemotron e DeepSeek, o Qwen foi avaliado nos três modos de prompt (few-shot, zero-shot e chain-of-thought).

3.2.3 Engenharia de Prompt

Para o modelo Big Pickle, três estratégias de prompt foram avaliadas, inspirando-se na taxonomia de Wei et al. [28]. O mesmo conjunto de três modos (few-shot, zero-shot e chain-of-thought) foi aplicado também ao Nemotron 3 Ultra, ao DeepSeek V4 Flash e ao Qwen 2.5 7B. O GPT-5.4 Mini, o GPT-5.4 (completo) e o Gemini 3.5 Flash foram avaliados apenas no modo Few-Shot por serem serviços pagos via OpenRouter; a investigação dos efeitos dos modos de prompt foi reservada aos modelos sem custo por chamada (Big Pickle, Nemotron, DeepSeek) e ao Qwen 2.5 7B (custo negligível). O prompt base compartilha as mesmas instruções gerais entre as variantes, diferindo na presença ou ausência de exemplos e na estrutura de raciocínio. A Tabela 9 detalha a estrutura de cada modo.

Tabela 9: Estrutura dos prompts nos três modos de simplificação.

Modo	Template do prompt (mensagens System + User)

Zero-Shot	**System:** ‘‘Você é um especialista em simplificação de textos jurídicos para Linguagem Simples. Seu papel é transformar decisões judiciais complexas em textos claros e acessíveis ao cidadão comum.’’
	User: `Texto jurídico original:` *[decisão original]*
	`Instruções:`
	1. Substitua todo o jargão jurídico por palavras simples.
	2. Divida frases longas em frases mais curtas.
	3. Mantenha todas as informações importantes: prazos, valores, nomes de artigos de lei.
	`Texto simplificado:`

Few-Shot	**System:** (mesmo) **User:** `Exemplo 1:` *Original: ‘‘In casu...’’* → *Simplificado: ‘‘Neste caso...’’* `Exemplo 2:` *Original: ‘‘O segurado postula...’’* → *Simplificado: ‘‘O trabalhador pediu...’’* `Exemplo 3:` *Original: ‘‘Trata-se de ação de divórcio...’’* → *Simplificado: ‘‘Este é um processo de divórcio...’’* `Texto jurídico original:` *[decisão original]* `Instruções:` (mesmos 3 itens) `Texto simplificado:`

CoT	**System:** (mesmo)
	User: (3 exemplos) + `Texto jurídico original:` *[decisão original]* + Instruções
	`Vamos pensar passo a passo:`
	1. Identifique quem são as partes envolvidas (quem pede e quem é requerido).
	2. Identifique qual é a decisão principal (o que o juiz decidiu).
	3. Identifique os detalhes importantes: valores, prazos, artigos de lei mencionados.
	4. Reescreva o texto em linguagem simples, eliminando jargões e dividindo frases longas, mas mantendo todos os fatos importantes.
	`Texto simplificado:`

[8] https://openrouter.ai/google/gemini-3.5-flash
[9] https://openrouter.ai/qwen/qwen-2.5-7b-instruct

3.3 Métricas de Avaliação

As simplificações foram avaliadas sob três dimensões complementares, totalizando 13 métricas.

3.3.1 Legibilidade

O **Índice Flesch Adaptado** para o português brasileiro [26] calcula a legibilidade com base na média de sílabas por palavra (S) e na média de palavras por sentença (P):

$$\text{Flesch} = 248 - 1,015P - 84,6S \tag{1}$$

A escala varia de 0 (extremamente difícil) a 100 (extremamente fácil). Para o português brasileiro, textos com pontuação abaixo de 40 são considerados de leitura muito difícil; entre 40 e 60, difíceis; entre 60 e 80, razoáveis; acima de 80, fáceis.

O **Gunning-Fog** [29] estima os anos de escolaridade necessários para compreensão do texto, considerando a média de palavras por sentença e o percentual de palavras complexas (três ou mais sílabas).

3.3.2 Similaridade Textual

O **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) [24] foi calculado nas variantes ROUGE-1 (unigramas), ROUGE-2 (bigramas) e ROUGE-L (maior subsequência comum). O ROUGE mede a sobreposição lexical entre o texto original e o simplificado.

3.3.3 Preservação Semântica

O **BERTScore** [25] utiliza embeddings contextuais para calcular similaridade semântica entre original e simplificado. Foram utilizados dois modelos: o multilingue `bert-base-multilingual-cased` [30] e o `neuralmind/bert-base-portuguese-cased` (BERTimbau) [23], específico para o português brasileiro. Para cada token do texto simplificado, o BERTScore computa a similaridade por cosseno com todos os tokens do original, selecionando a melhor correspondência. As métricas de precisão (P) e revocação (R) são combinadas no F1. Ambos os modelos foram calculados para efeito de comparação; as discussões ao longo do artigo referem-se ao mBERT, salvo indicação contrária.

Para eficiência computacional, os 100 pares de cada técnica foram processados em lote (batch) na GPU: a lista de pares é passada ao modelo em uma única chamada, reduzindo o overhead de carregamento, mas as métricas de precisão, revocação e F1 são computadas individualmente por par de documentos, preservando a granularidade necessária para a análise estatística pareada (Seção 4.7).

3.4 Protocolo Experimental

Cada uma das 100 decisões foi simplificada por todas as dezesseis técnicas, totalizando 1.700 simplificações (100 originais + 1.600 simplificadas). Para cada par (original, simplificado), foram calculadas 13 métricas: Flesch Adaptado, Gunning-Fog, ganho absoluto de Flesch (ΔFlesch), ROUGE-1, ROUGE-2, ROUGE-L, BERTScore (mBERT) precisão, BERTScore (mBERT) revocação, BERTScore (mBERT) F1, BERTScore (BERTimbau) precisão, BERTScore (BERTimbau) revocação, BERTScore (BERTimbau) F1 e variação de tamanho (palavras). As 100 decisões foram processadas na íntegra, sem truncamento. O tempo médio de resposta por documento foi de aproximadamente 30 segundos para Big Pickle, 40 segundos para Nemotron, 35 segundos para DeepSeek, 9 segundos para Qwen 2.5 7B, 2,5 segundos para GPT-5.4 Mini, 4 segundos para GPT-5.4 (completo) e 2 segundos para Gemini 3.5 Flash, totalizando cerca de 7 horas de processamento para geração completa do corpus. Todos os LLMs foram acessados entre 12 e 13 de junho de 2026, via API do OpenCode Zen e OpenRouter.

Para o cálculo do BERTScore, utilizaram-se os modelos multilingue `bert-base-multilingual-cased` [30] (camada 9, `idf=False`, `lang="pt"`) e o `neuralmind/bert-base-portuguese-cased` (BERTimbau) [23] (camada 12, última camada).

Ressalta-se que as tabelas de resultados (Tabelas 2 a 5) reportam as **médias** amostrais para compatibilidade com a literatura de simplificação textual. Os testes estatísticos (Seção 4.7), por sua vez, empregam **medianas** — mais robustas a outliers — e o teste não-paramétrico de Friedman, que opera sobre *ranks*. A escolha é motivada pela presença de outliers extremos em algumas variantes (e.g., Qwen 2.5 7B Zero-Shot apresentou um ΔFlesch de $-778,6$ em um documento, decorrente de um artefato de geração por repetição infinita), que distorceriam a média mas não afetam a mediana nem os testes baseados em *ranks*. O documento com repetição infinita foi identificado por inspeção (tamanho superior a 50.000 caracteres) e re-gerado com o mesmo prompt e temperatura 0,3. A nova saída (394 palavras, ΔFlesch = $+13,2$) substituiu a anterior no cálculo das médias. Sem esta correção, a média do Qwen 2.5 7B Zero-Shot seria de $+5,4$; com a correção, passou para $+13,6$, valor próximo à mediana de $+13,0$.

Adicionalmente, realizou-se uma análise qualitativa de erros via *LLM-as-judge*: para cada par (original, simplificado) de cada variante LLM, um modelo juiz (GPT-5.4 Mini) classificou a saída em cinco categorias de erro (alucinação de valores, invenção processual, generalização excessiva, perda de nuances e falha de formatação), totalizando 1.500 avaliações adicionais. Os resultados são discutidos na Seção 4.6. A Tabela 10 apresenta o prompt utilizado.

Tabela 10: Prompt utilizado pelo LLM-as-Judge (GPT-5.4 Mini) para classificação de erros nas simplificações.

Componente	**Conteúdo**
System	‘‘Você é um sistema classificador de erros de PLN especializado no domínio jurídico brasileiro. Sua tarefa é analisar um par de documentos (Texto Original vs. Texto Simplificado) e determinar, categoricamente, a presença ou ausência de falhas de simplificação.’’
Categorias	**[ALUCINACAO_VALORES]**: ocorreu se o texto simplificado introduziu, alterou ou distorceu prazos, valores monetários, idades ou porcentagens que não estavam no texto original. **[INVENCAO_PROCESSUAL]**: ocorreu se o texto simplificado inventou andamentos, datas de julgamento ou resultados do recurso que não constam no original. **[GENERALIZACAO_EXCESSIVA]**: ocorreu se o texto substituiu referências fundamentais (como artigos de lei ou nomes de órgãos) por termos vagos demais. **[PERDA_NUANCES]**: ocorreu se o texto omitiu condições jurídicas ou exceções essenciais para o entendimento do caso. **[FORMATACAO_INCONSISTENTE]**: ocorreu se o texto simplificado misturou de forma desordenada parágrafos, listas e estruturas de perguntas e respostas.
Instruções	Responda completa e exclusivamente com um objeto JSON válido. Se a categoria ocorreu, defina `identificado` como `true` e forneça o trecho do texto simplificado onde o erro aparece em `evidencia`. Se não ocorreu, defina `identificado` como `false` e `evidencia` como `null`.
Formato JSON	`{ ``id_documento'': ``[ID]'', ``erros'': { ``alucinacao_valores'': { ``identificado'': true/false, ``evidencia'': ``string ou null'' }, ``invencao_processual'': { ... }, ``generalizacao_excessiva'': { ... }, ``perda_nuances'': { ... }, ``falha_formatacao'': { ... } } }`

4 Resultados

4.1 Visão Geral

A Tabela 2 apresenta os resultados comparativos das dezesseis técnicas de simplificação.

A abordagem baseada em regras apresentou ganho negativo de $-1,6$ pontos no índice Flesch, reduzindo a legibilidade de 39,7 para 38,1. Este resultado confirma que as substituições lexicais isoladas são

insuficientes para melhorar a legibilidade — e, em alguns casos, podem até piorá-la ao introduzir termos que alteram a fluência do texto sem modificar sua estrutura sintática complexa.

Em contraste, todas as abordagens baseadas em LLMs obtiveram ganhos, embora com variações importantes. O **DeepSeek V4 Flash** e o **Big Pickle (Few-Shot)** alcançaram os maiores ganhos (+24,3 pontos cada), elevando o Flesch médio para 64,1 e 64,0, respectivamente (razoável), seguidos pelo Nemotron 3 Ultra (+22,7). O **Qwen 2.5 7B (CoT)** (+7,2) apresentou o menor ganho entre os LLMs; o modo Zero-Shot alcançou +13,6 e o Few-Shot +13,9. O Qwen 2.5 7B (Zero-Shot) obteve a maior preservação semântica (BERTScore mBERT de 0,748), evidenciando o trade-off entre legibilidade e fidelidade ao texto original — modelos menores tendem a alterar menos a estrutura sintática, o que reduz o ganho de legibilidade mas preserva melhor o significado. O **GPT-5.4 (completo)** (+17,6) e o **GPT-5.4 Mini** (+16,4) ocupam uma posição intermediária nesse espectro. Para referência, [18] discutem ganhos de legibilidade obtidos com o uso de linguagem simples em contratos jurídicos em inglês, com resultados consistentes com os observados em nosso estudo, embora as diferenças de idioma e a maior diversidade do nosso corpus devam ser consideradas na comparação.

Em termos de preservação semântica (BERTScore F1, mBERT), o **Qwen 2.5 7B (Zero-Shot)** lidera com 0,748, superando o GPT-5.4 (completo) (0,713) e o Nemotron 3 Ultra (0,704). O ROUGE-1 é especialmente alto para o GPT-5.4 (completo) (0,583), sugerindo que sua abordagem mais conservadora preserva substancialmente o vocabulário e a estrutura lexical do original — embora o Qwen 2.5 7B, com apenas 7 bilhões de parâmetros, alcance valor comparável (0,483) a um custo computacional muito menor.

A análise do tamanho dos textos revela outra dimensão importante da simplificação. O texto original médio tem 527 palavras. As simplificações do Big Pickle few-shot reduzem para 249 palavras (média de 53% de compressão), enquanto o DeepSeek produz textos de 255 palavras (52% de compressão) e o Nemotron, 271 palavras (49% de compressão). O GPT-5.4 Mini produz textos de 218 palavras (59% de compressão), alinhado com seu perfil mais conservador. O GPT-5.4 (completo), por outro lado, praticamente mantém o tamanho original (438 palavras, compressão de 17%), indicando que sua estratégia de simplificação prioriza substituições lexicais pontuais em vez de reestruturação sintática agressiva. A técnica baseada em regras também mantém o tamanho original (517 palavras, compressão de apenas 2%), consistente com sua abordagem conservadora de substituições lexicais pontuais. O Gemini 3.5 Flash produziu os textos mais concisos (120 palavras, 77% de compressão), consistente com seu perfil de baixa latência. Essa alta taxa de compressão explica seu ROUGE-1 comparativamente baixo (0,255, o menor entre todas as técnicas): textos muito concisos necessariamente apresentam menor sobreposição lexical com o original, o que não reflete necessariamente perda de qualidade semântica. O modo CoT do Big Pickle produziu textos de 225 palavras (57% de compressão), sugerindo que o modelo, ao ‘‘explicar’’ a decisão, tende a resumir em vez de simplificar — padrão também observado no Nemotron CoT, DeepSeek CoT e Qwen CoT.

A Figura 2 ilustra visualmente o ganho de legibilidade por técnica, evidenciando a lacuna entre as abordagens baseadas em LLMs e a baseline de regras.

4.2 Resultados por Tema

A Tabela 3 revela variações significativas entre os temas jurídicos.

Para o tema **consumidor**, o DeepSeek V4 Flash e o Big Pickle few-shot obtiveram os maiores ganhos (+28,2 e +27,4 pontos), seguidos pelo Nemotron (+25,5). Uma possível explicação para o bom desempenho no tema consumidor é a natureza mais concreta e cotidiana das disputas (negativa de cobertura de plano de saúde, defeitos em produtos), que facilitam a paráfrase pelo LLM.

Para o tema **previdenciário**, os ganhos foram mais moderados que o esperado: DeepSeek (+23,0), Big Pickle (+22,5) e Nemotron (+20,9). Embora os textos previdenciários originais tenham alta complexidade (Flesch médio de 40,8, classificado como ‘‘difícil’’), a variação entre documentos é grande — alguns documentos apresentam ganhos elevados enquanto outros, ganhos modestos — resultando em médias agregadas mais contidas que os máximos individuais.

O **GPT-5.4 (completo)** e o **GPT-5.4 Mini** seguiram o mesmo padrão temático, com ganhos mais conservadores: consumidor (+19,9 e +19,5), previdenciário (+16,4 e +12,4) e família (+16,3 e +17,1),

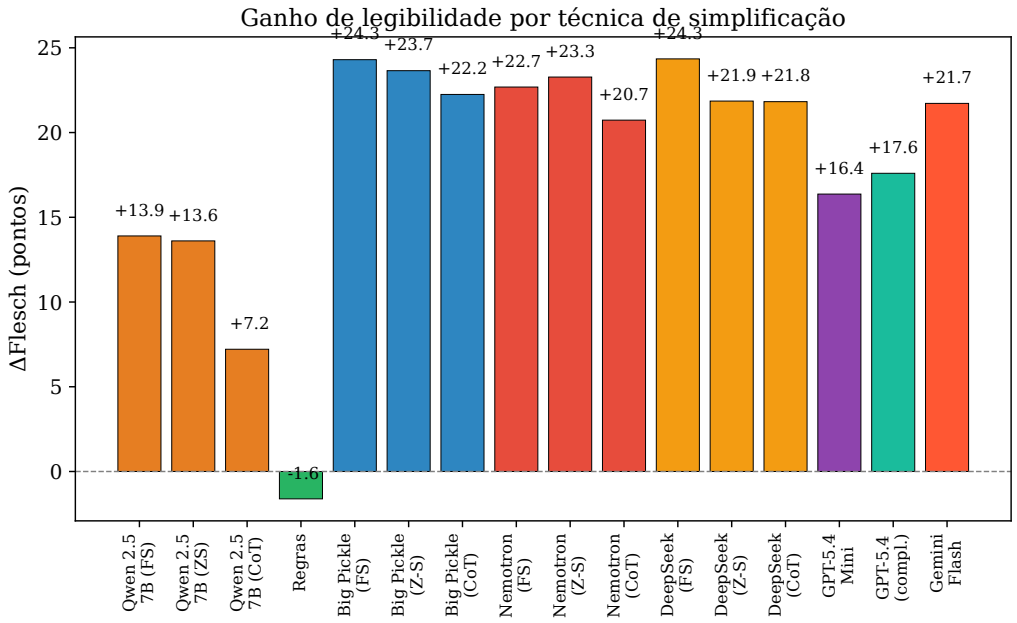


Figura 2: Ganho de legibilidade (Flesch Adaptado) por técnica de simplificação.

confirmando que sua abordagem menos intervencionista é consistente entre domínios.

O tema **família** mostrou-se o mais desafiador: entre os modelos de médio e grande porte, os ganhos variaram de +21,6 (Nemotron) a +22,9 (Big Pickle few-shot); o Qwen 2.5 7B obteve ganhos expressivamente menores nesse tema (+8,4 FS, +12,1 ZS, +7,5 CoT), sugerindo que modelos menores têm dificuldade adicional com a terminologia sensível do direito de família. Os resultados sugerem que questões familiares envolvem terminologia mais sensível e contextual (guarda compartilhada, alienação parental, regime de bens), onde o modelo precisa equilibrar simplificação com precisão jurídica e emocional.

A Tabela 5 apresenta o índice Gunning-Fog absoluto e a variação percentual por tema para todas as técnicas. As regras produzem ligeiro aumento do Fog (+0,7–0,9%), enquanto todos os LLMs o reduzem (7,6–21,9%), com destaque para o Nemotron 3 Ultra (Few-Shot) nos temas Família (17,3) e Previdenciário (17,2) e para o Nemotron 3 Ultra (Zero-Shot) em Consumidor (17,4).[10] Confirmando o padrão observado no ganho Flesch, o Nemotron e o GPT-5.4 (completo) produzem as maiores reduções (17,2–18,8), enquanto o GPT-5.4 Mini é o LLM mais conservador (19,0–19,5). A técnica baseada em regras, consistente com o ganho negativo de Flesch, eleva ligeiramente o Fog em relação ao original, indicando que substituições lexicais isoladas não reduzem a complexidade sintática necessária para baixar o índice.

O Gemini 3.5 Flash, apesar de produzir os textos mais concisos (77% de compressão), apresentou redução de Fog modesta em previdenciário (-8,1%), sugerindo que sua compressão agressiva retém palavras jurídicas curtas porém densas (ex.: INSS, BPC, LOAS) que elevam o índice Fog sem comprometer a legibilidade — capturada pelo ganho Flesch de +18,0 no mesmo tema.

4.3 Comparação dos Modos de Prompt

A Tabela 5 também detalha esses valores por tema e modo de prompt.

No Big Pickle, o modo **CoT** apresentou o pior desempenho entre suas variantes: menor ganho Flesch (+22,2), menor BERTScore F1 (0,655) e menor ROUGE-1 (0,336). O mesmo padrão foi observado nos demais modelos — Qwen CoT (+7,2), Nemotron CoT (+20,7) e DeepSeek CoT (+21,8) — confirmando que o CoT é sistematicamente inferior ao Few-Shot e ao Zero-Shot para esta tarefa. Este resultado contraria a expectativa inicial, baseada em Wei et al. [28], de que o raciocínio passo a passo melhoraria a qualidade da simplificação. Uma análise qualitativa das saídas CoT revela que o modelo frequentemente produz

[10] Valores negativos na coluna Variação indicam redução da complexidade textual.

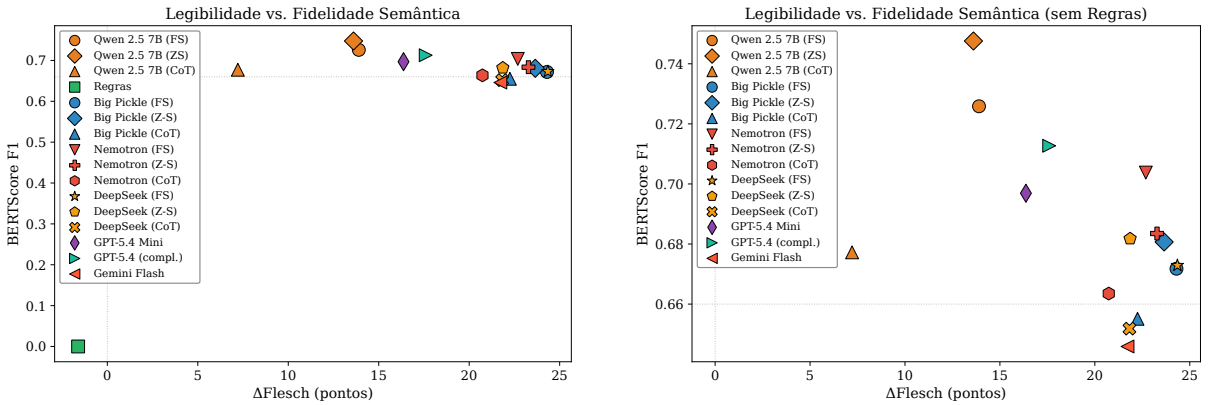
explicações genéricas sobre a decisão (‘‘o tribunal decidiu sobre...’’) em vez de simplificar o texto original propriamente dito. No domínio jurídico, as etapas genéricas de raciocínio do CoT (‘‘identifique as partes’’) não capturam adequadamente as especificidades terminológicas e estruturais dos textos legais. O modo **zero-shot** apresentou ganho de +23,7 e ROUGE-1 de 0,433, superando o few-shot em similaridade lexical, mas com menor ganho de legibilidade. O modo **few-shot** lidera em legibilidade (+24,3), demonstrando o valor do aprendizado por analogia para esta tarefa.

4.4 BERTScore e Análise de Trade-offs

A Tabela 4 detalha os resultados completos do BERTScore, enquanto a Figura 3 visualiza a relação entre ganho de legibilidade e preservação semântica.

A técnica baseada em regras apresenta BERTScore F1 de 0,994, o que aparentemente contradiz seu ganho negativo de legibilidade. Isto ocorre porque as regras alteram minimamente o texto original (substituições lexicais pontuais), resultando em alta similaridade superficial, mas sem efetivamente simplificar a estrutura sintática. Este fenômeno ilustra a importância de utilizar métricas complementares: ROUGE e BERTScore, isoladamente, não capturam a melhoria de legibilidade.

A Figura 3 revela uma fronteira de Pareto entre legibilidade e fidelidade semântica. No extremo esquerdo inferior, a técnica baseada em regras (baixo ganho, baixa legibilidade efetiva). No extremo direito, Big Pickle few-shot (máxima legibilidade, BERTScore moderado). O Qwen 2.5 7B (Zero-Shot) alcança o BERTScore máximo (0,748) com ganho modesto de legibilidade (+13,6); o GPT-5.4 (completo) ocupa o segundo lugar (0,713, +17,6). O **GPT-5.4 Mini** ocupa um quadrante próximo: alto BERTScore (0,697) com ganho moderado (+16,4). O Nemotron (0,704, +22,7) oferece o melhor equilíbrio entre as duas dimensões entre os modelos de maior ganho. A escolha ideal depende do contexto de uso: para cidadãos com baixo letramento, prioriza-se legibilidade; para operadores do direito, prioriza-se fidelidade semântica.



(a) Todas as técnicas

(b) Sem técnica baseada em regras

Figura 3: Relação entre ganho de legibilidade (Flesch) e preservação semântica (BERTScore F1) para cada técnica.

4.5 Análise Qualitativa

Para complementar as métricas quantitativas, realizamos uma análise qualitativa comparando as saídas de diferentes técnicas sobre a mesma decisão original.

As Tabelas 6, 7 e 8 apresentam três exemplos representativos, um de cada tema, com as simplificações das principais técnicas. A seguir, discutimos esses e outros exemplos adicionalmente no texto. Por limitação de espaço, são apresentados na tabela apenas os modos Few-Shot para os modelos com múltiplos prompts; os padrões dos modos Zero-Shot e CoT são discutidos na Seção 4.3.

Exemplo 1 (Previdenciário — RE 626837): O texto original discute a imunidade recíproca para contribuição previdenciária de agentes políticos. O Big Pickle few-shot produziu uma explicação clara e estruturada: ‘‘Decisão do STF sobre INSS de políticos (Tema 691). O Supremo Tribunal Federal (STF) julgou um recurso sobre a cobrança do INSS para políticos eleitos (prefeitos, vereadores, deputados, etc.)’’. A técnica baseada em regras, por outro lado, limitou-se a substituir alguns termos (‘‘in casu’’ por ‘‘neste caso’’), mantendo a estrutura complexa original. O DeepSeek produziu uma versão que explica de forma direta que ‘‘a Prefeitura não precisa pagar o INSS patronal sobre os salários dos políticos’’ mas omitiu a tese de repercussão geral, diferentemente do Big Pickle few-shot que a preservou integralmente.

Exemplo 2 (Consumidor — RE 1499539 RG): A ementa discute a reestruturação de carreira de magistério municipal e a paridade remuneratória. O Big Pickle simplificou para: ‘‘Um professor entrou com recurso argumentando que seu salário deveria ser igual ao dos professores da ativa. O tribunal de origem negou o pedido sob o argumento de que a paridade só vale para quem se aposentou até 2003’’. O Nemotron preservou mais a estrutura formal, enquanto o DeepSeek produziu uma simplificação intermediária.

Exemplo 3 (Família — RE 842844): A decisão trata dos direitos de servidora pública gestante em cargo comissionado à licença-maternidade e estabilidade provisória. O Big Pickle few-shot produziu: ‘‘O STF decidiu que toda servidora pública grávida tem direito à licença-maternidade e à estabilidade provisória, mesmo que ela não seja concursada. Isso vale para quem ocupa cargo de confiança (comissionado) ou tem contrato temporário com o governo’’. O DeepSeek acrescentou novos exemplos concretos: ‘‘se a servidora é contratada por tempo determinado, ainda assim ela não pode ser dispensada durante a gravidez e até 5 meses após o parto’’, detalhando o período de estabilidade, informação que estava implícita no original. O Nemotron produziu uma resposta mais formal, organizando a decisão em tópicos numerados, útil para operadores do direito, mas potencialmente intimidadora para o cidadão comum. A técnica baseada em regras, mais uma vez, não conseguiu ir além de substituições lexicais, mantendo inalterada a estrutura complexa da ementa.

Exemplo 4 (Previdenciário — RE 1412093): O texto original discute a possibilidade de acumular benefícios previdenciários com o Benefício de Prestação Continuada (BPC/LOAS). O Big Pickle few-shot explicou que ‘‘quem já recebe aposentadoria ou pensão por morte pode também receber o BPC/LOAS’’, mas alertou que ‘‘os valores somados não podem ultrapassar 1/4 do salário mínimo por pessoa da família’’. O DeepSeek produziu uma explicação estruturada em perguntas e respostas: ‘‘Quem tem direito? Pessoas com deficiência e idosos com 65 anos ou mais, de qualquer idade, que comprovem baixa renda’’, enquanto o Nemotron manteve a estrutura de tópicos, com linguagem precisa porém formal.

Exemplo 5 (Consumidor — Tema 745 ICMS Seletividade): O STF decidiu sobre a impossibilidade de estados cobrarem ICMS maior sobre energia elétrica e telecomunicações. O Big Pickle few-shot explicou: ‘‘Se o seu estado escolheu cobrar alíquotas diferentes de ICMS, ele não pode cobrar um imposto maior sobre a energia elétrica e os serviços de telecomunicação do que a alíquota cobrada sobre a maioria dos outros produtos’’, adicionando uma explicação temporal sobre a modulação de efeitos. O Nemotron, por sua vez, produziu uma resposta mais técnica, listando explicitamente os artigos constitucionais e a fundamentação jurídica, com menor preocupação com a acessibilidade imediata, mas maior precisão terminológica.

O **GPT-5.4 (completo)** e o **GPT-5.4 Mini** apresentaram estilos consistentes de simplificação moderada nos exemplos analisados. O GPT-5.4 Mini mantém a estrutura argumentativa original e substitui termos jurídicos, mas preserva mais a sintaxe original que os demais LLMs. Por exemplo, no RE 842844 (família), sua saída manteve a organização em parágrafos da ementa original, diferentemente da reestruturação mais agressiva do Big Pickle ou da formatação em tópicos do Nemotron. O GPT-5.4 (completo), por sua vez, produz textos mais longos e preserva substancialmente o vocabulário original, resultando no maior ROUGE-1 (0,583) e no segundo maior BERTScore (0,713) entre todos os LLMs, mas com menor redução de tamanho (∼17% de compressão).

4.6 Análise de Erros

A análise qualitativa também identificou padrões de erro recorrentes nas simplificações geradas por LLMs. Para sistematizar essa análise, utilizamos o *GPT-5.4 Mini* como juiz automático (*LLM-as-judge*), avaliando as 15 variantes de LLM sobre o corpus completo de 100 documentos (1500 avaliações no total). O prompt do juiz apresenta o par (original, simplificado) e instrui o modelo a classificar cada uma das cinco categorias de erro, fornecendo a evidência textual em formato JSON obrigatório, sem raciocínio intermediário. As categorias avaliadas foram: alucinação de valores (prazos, percentuais inexistentes), invenção de informações processuais (datas, andamentos fictícios), generalização excessiva (substituição de referências normativas precisas por termos vagos), perda de nuances (omissão de condições ou exceções jurídicas essenciais) e falha de formatação (mistura inconsistente de estruturas textuais). A Tabela 11 apresenta a frequência de cada categoria de erro por técnica:

- **Alucinação de valores**: LLMs introduzem valores numéricos incorretos (prazos, percentuais) que não constam no original. O GPT-5.4 (completo) apresenta a menor taxa (7/100), enquanto o Qwen 2.5 7B few-shot alcança 49/100, indicando forte correlação entre capacidade do modelo e fidelidade factual. O modo CoT reduz alucinações nos modelos menores (Qwen: 49→37; DeepSeek: 34→24), mas tem efeito misto nos maiores;
- **Invenção de informações processuais**: casos em que o simplificado acrescenta dados processuais fictícios (datas, números de processos, nomes de juízes). GPT-5.4 (completo) novamente lidera (3/100), seguido por Gemini 3.5 Flash (13/100) e Nemotron 3 Ultra zero-shot (15/100). O modo zero-shot consistentemente reduz esse erro em relação ao few-shot para a maioria dos modelos, possivelmente porque o few-shot fornece exemplos que o modelo tenta imitar excessivamente;
- **Generalização excessiva**: simplificações que omitem detalhes jurídicos relevantes, substituindo referências precisas (artigos de lei, Súmulas) por expressões genéricas como ‘‘a lei diz’’ ou ‘‘segundo a Constituição’’. As taxas variam de 8/100 (GPT-5.4 completo) a 57/100 (Qwen 2.5 7B few-shot). Modelos maiores preservam melhor a referência normativa específica;
- **Perda de nuances**: o erro mais frequente em todas as técnicas, afetando de 34/100 (GPT-5.4 completo) a 80/100 (Qwen 2.5 7B few-shot). Decisões que envolvem distinções sutis entre conceitos jurídicos (ex.: ‘‘aposentadoria por tempo de contribuição com pedágio’’ vs. ‘‘aposentadoria especial’’) são simplificadas de forma genérica. Este padrão reflete uma limitação fundamental da simplificação automática: o trade-off entre acessibilidade e completude informacional;
- **Falha de formatação**: mistura inconsistente de tópicos, parágrafos e listas entre documentos do mesmo modelo. DeepSeek V4 Flash CoT apresenta a menor taxa (23/100), indicando que o modo corrente-de-pensamento favorece estruturação consistente. Surpreendentemente, GPT-5.4 (completo) também apresenta uma das piores taxas (57/100), sugerindo que seu texto mais longo e preservativo do original também retém a formatação heterogênea das ementas;
- **Correlação com tamanho do modelo**: observa-se uma correlação inversa consistente entre capacidade do modelo e taxa de erros. GPT-5.4 (completo) lidera em 4 das 5 categorias; GPT-5.4 Mini e Gemini 3.5 Flash ocupam posições intermediárias (a baixa taxa de alucinação do Gemini deve-se em parte à sua alta compressão de 77%, que reduz o número de proposições geradas e, consequentemente, as oportunidades de erro factual); Qwen 2.5 7B (o menor modelo avaliado) apresenta as maiores taxas na maioria das categorias. Modelos médios (Nemotron 3 Ultra, DeepSeek V4 Flash) situam-se entre os extremos, com vantagens específicas em categorias particulares (DeepSeek CoT em formatação; Nemotron zero-shot em generalização excessiva).

Registra-se que o GPT-5.4 Mini, utilizado como juiz, é também uma das técnicas avaliadas, o que pode introduzir viés de familiaridade com o próprio estilo de saída — por exemplo, maior tolerância a erros de formatação característicos de sua própria geração ou maior severidade com padrões estruturais divergentes. Estudos futuros devem replicar a análise com um juiz de família de modelo distinta para verificar a robustez dos resultados aqui reportados.

A técnica baseada em regras, embora não apresente nenhum desses tipos de alucinação, falha em melhorar a legibilidade por não reestruturar as sentenças. Este trade-off fundamental entre segurança

Tabela 11: Matriz de erros por técnica de simplificação (N=100 documentos, ≈33 por tema)

Técnica	Alucinação	Inv. Processual	Gen. Excessiva	Perda Nuances	Falha Format.
Qwen 2.5 7B (FS)	49	45	57	80	63
Qwen 2.5 7B (ZS)	45	29	49	75	59
Qwen 2.5 7B (CoT)	37	31	40	64	48
Big Pickle (FS)	34	24	35	61	47
Big Pickle (ZS)	28	17	33	69	58
Big Pickle (CoT)	33	27	43	67	35
Nemotron 3 Ultra (FS)	34	19	30	60	57
Nemotron 3 Ultra (ZS)	24	15	28	51	53
Nemotron 3 Ultra (CoT)	31	23	34	62	39
DeepSeek V4 Flash (FS)	34	25	41	63	47
DeepSeek V4 Flash (ZS)	29	20	38	59	55
DeepSeek V4 Flash (CoT)	24	23	41	60	23
GPT-5.4 Mini	19	18	21	47	37
GPT-5.4 (completo)	7	3	8	34	57
Gemini 3.5 Flash	18	13	28	66	43

(regras) e eficácia (LLMs) deve ser considerado no design de sistemas reais de simplificação jurídica.

4.7 Testes Estatísticos

Para verificar a significância estatística das diferenças observadas entre as técnicas, realizamos três testes complementares: Friedman (teste global não-paramétrico), Nemenyi (pós-teste pareado) e Wilcoxon pareado com correção de Bonferroni ($\alpha_{Bonf} = 0,05/120 \approx 0,000417$ para 16 técnicas; $0,05/105 \approx 0,000476$ para BERTScore, que exclui a baseline de regras). A Tabela 12 apresenta os resultados completos.

O teste de Friedman rejeita a hipótese nula de que todas as técnicas têm a mesma distribuição de ganho para as três métricas ($p < 0,001$ em todos os casos). O pós-teste de Nemenyi confirma que as abordagens baseadas em LLMs diferem significativamente da baseline de regras ($p < 0,05$). Entre os LLMs, as diferenças são mais pronunciadas com a inclusão dos modos adicionais de prompt: no ganho Flesch, boa parte dos 120 pares comparados são significativos pelo teste de Wilcoxon com correção de Bonferroni, refletindo a diversidade entre as 16 técnicas.

Para o ROUGE-1, a maioria dos pares é significativa, indicando que as técnicas produzem sobreposições lexicais distintas. O BERTScore F1 apresenta diferenças significativas na maioria dos pares entre LLMs, refletindo variações na preservação semântica. A Tabela 13 detalha as comparações pareadas, e a Tabela 14 apresenta as estatísticas descritivas do ganho Flesch. Em complemento aos p-valores, calculou-se o d de Cohen pareado para pares representativos: Big Pickle FS vs. Regras apresentou d = 1,83 (magnitude grande), DeepSeek V4 Flash FS vs. Qwen 2.5 7B FS apresentou d = 1,21 (grande), enquanto GPT-5.4 Mini vs. GPT-5.4 (completo) apresentou d = 0,31 (pequeno), confirmando que a significância estatística detectada reflete magnitudes distintas de efeito.

As Figuras 4, 5 e 6 apresentam os diagramas de diferenças críticas de Nemenyi.

Tabela 12: Testes estatísticos das métricas de simplificação. Friedman avalia diferença global entre técnicas; Nemenyi e Wilcoxon (correção de Bonferroni) identificam pares com diferença significativa.

Métrica	Friedman χ^2	p	Wilcoxon signif.	Nemenyi $p < 0,05$	Total pares
Ganho Flesch	561.12	$< 0.001^{***}$	71/120	65/120	120
ROUGE-1	839.15	$< 0.001^{***}$	93/120	81/120	120
BERTScore F1	601.25	$< 0.001^{***}$	80/105	70/105	105

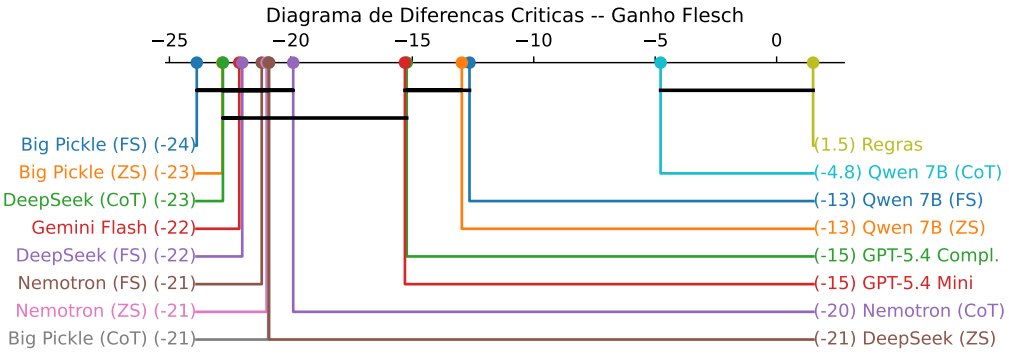


Figura 4: Diagrama de diferenças críticas — Ganho Flesch.

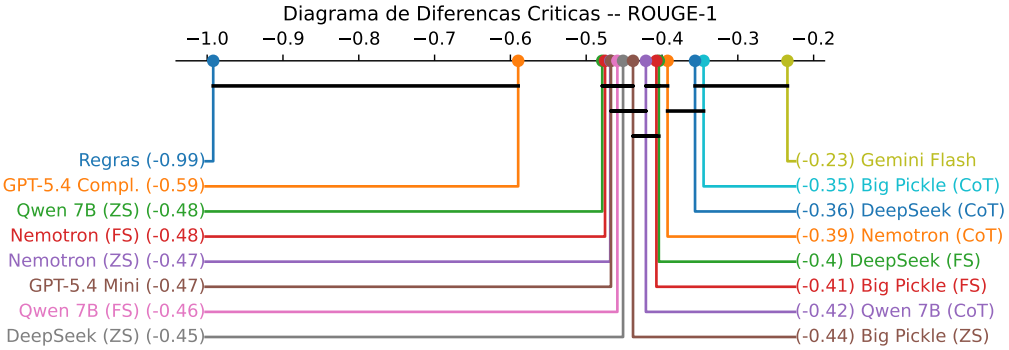


Figura 5: Diagrama de diferenças críticas — ROUGE-1.

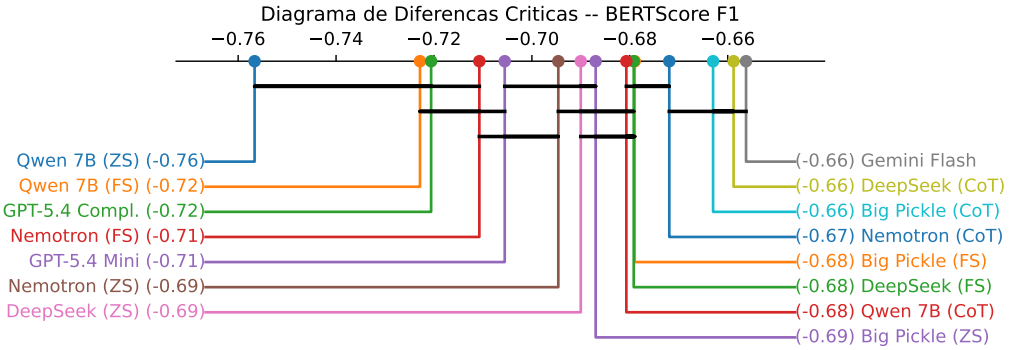


Figura 6: Diagrama de diferenças críticas — BERTScore F1.

Tabela 13: Medianas de ΔFlesch e ROUGE-1 e Média Amostral do BERTScore F1 por técnica, com resultados do teste de Wilcoxon pareado com correção de Bonferroni ($\alpha_{Bonf} = 0,05/120 \approx 0.000417$ para Flesch e ROUGE; $0,05/105 \approx 0.000476$ para BERTScore). Testes significativos indicados com $p < 0,05$, $p < 0,01$, $p < 0,001$.

Técnica	ΔFlesch (pts)	ROUGE-1	BS F1	vs Regras	vs Big Pickle (FS)	vs GPT-5.4 Mini
Qwen 2.5 7B (Few-Shot)	+12.6	0.4589	0.7229	$< 0.001^{***}$	$< 0.001^{***}$	0.065
Qwen 2.5 7B (Zero-Shot)	+13.0	0.4788	0.7566	$< 0.001^{***}$	$< 0.001^{***}$	0.018^{*}
Qwen 2.5 7B (CoT)	+4.8	0.4213	0.6807	$< 0.001^{***}$	$< 0.001^{***}$	$< 0.001^{***}$
Big Pickle (Few-Shot)	+23.9	0.4074	0.6790	$< 0.001^{***}$	—	$< 0.001^{***}$
Big Pickle (Zero-Shot)	+22.8	0.4382	0.6870	$< 0.001^{***}$	0.314	$< 0.001^{***}$
Big Pickle (CoT)	+20.9	0.3451	0.6630	$< 0.001^{***}$	0.053	$< 0.001^{***}$
Nemotron 3 Ultra (Few-Shot)	+21.2	0.4752	0.7107	$< 0.001^{***}$	0.206	$< 0.001^{***}$
Nemotron 3 Ultra (Zero-Shot)	+21.0	0.4683	0.6946	$< 0.001^{***}$	0.315	$< 0.001^{***}$
Nemotron 3 Ultra (CoT)	+19.9	0.3926	0.6719	$< 0.001^{***}$	$< 0.001^{***}$	$< 0.001^{***}$
DeepSeek V4 Flash (Few-Shot)	+22.0	0.4042	0.6792	$< 0.001^{***}$	0.918	$< 0.001^{***}$
DeepSeek V4 Flash (Zero-Shot)	+20.9	0.4514	0.6900	$< 0.001^{***}$	0.004^{**}	$< 0.001^{***}$
DeepSeek V4 Flash (CoT)	+22.8	0.3564	0.6588	$< 0.001^{***}$	0.011^{*}	$< 0.001^{***}$
GPT-5.4 Mini	+15.3	0.4673	0.7055	$< 0.001^{***}$	$< 0.001^{***}$	—
GPT-5.4 (completo)	+15.2	0.5897	0.7206	$< 0.001^{***}$	$< 0.001^{***}$	0.227
Gemini 3.5 Flash	+22.1	0.2345	0.6563	$< 0.001^{***}$	0.013^{*}	$< 0.001^{***}$
Baseado em Regras	-1.5	0.9920	—	—	$< 0.001^{***}$	$< 0.001^{***}$

Tabela 14: Estatísticas descritivas do ganho Flesch (pontos) por técnica.

Técnica	Média	Mediana	DP	Min.	Max.	IQ
Qwen 2.5 7B (Few-Shot)	+13.9	+12.6	14.1	-11.7	+60.4	17.8
Qwen 2.5 7B (Zero-Shot)	+13.6	+13.0	14.6	-10.2	+69.5	19.8
Qwen 2.5 7B (CoT)	+7.2	+4.8	14.7	-23.5	+56.7	15.7
Big Pickle (Few-Shot)	+24.3	+23.9	14.5	-3.6	+92.6	15.0
Big Pickle (Zero-Shot)	+23.7	+22.8	14.5	-5.6	+77.7	18.7
Big Pickle (CoT)	+22.2	+20.9	16.2	-11.4	+97.4	18.8
Nemotron 3 Ultra (Few-Shot)	+22.7	+21.2	13.2	-7.4	+91.0	14.2
Nemotron 3 Ultra (Zero-Shot)	+23.3	+21.0	14.0	-0.5	+101.3	14.0
Nemotron 3 Ultra (CoT)	+20.7	+19.9	14.7	-4.3	+85.4	14.3
DeepSeek V4 Flash (Few-Shot)	+24.3	+22.0	15.1	+2.3	+93.9	18.7
DeepSeek V4 Flash (Zero-Shot)	+21.9	+20.9	12.7	-2.3	+72.5	14.6
DeepSeek V4 Flash (CoT)	+21.8	+22.8	16.1	-12.0	+74.3	20.5
GPT-5.4 Mini	+16.4	+15.3	12.6	-12.3	+65.6	14.1
GPT-5.4 (completo)	+17.6	+15.2	11.9	+0.8	+86.6	11.2
Gemini 3.5 Flash	+21.7	+22.1	16.0	-11.6	+78.3	19.9
Baseado em Regras	-1.6	-1.5	1.1	-6.5	+0.0	1.2

5 Discussão

Os resultados obtidos permitem extrair várias lições importantes para a simplificação automática de textos jurídicos no Brasil. Organizamos a discussão em sete eixos: (i) a comparação entre abordagens neurais e simbólicas, (ii) as limitações das métricas automáticas, (iii) os efeitos das estratégias de prompt, (iv) a comparação com resultados da literatura, (v) as implicações práticas para o Judiciário, (vi) as considerações éticas envolvidas, e (vii) as ameaças à validade do estudo.

5.1 LLMs vs. Regras

A superioridade consistente dos LLMs sobre a abordagem baseada em regras confirma que a simplificação textual jurídica exige mais do que substituições lexicais. Os LLMs mais agressivos (DeepSeek, Big Pickle, Nemotron) obtiveram ganhos entre +21 e +24 pontos, enquanto o GPT-5.4 (completo) e o GPT-5.4 Mini, mais conservadores, alcançaram +17,6 e +16,4 — ainda assim muito superior ao desempenho negativo das regras ($-1,6$). As regras não conseguem endereçar a complexidade sintática das decisões judiciais, que frequentemente apresentam:

- Orações subordinadas encadeadas (até 4 níveis de profundidade);
- Intercalações explicativas entre sujeito e verbo;
- Estrutura argumentativa densa com citações jurisprudenciais e doutrinárias;
- Períodos que excedem 50 palavras.

Os LLMs, ao compreenderem o texto em nível semântico, conseguem reestruturá-lo mantendo a cadeia argumentativa, o que é fundamental para a utilidade prática da simplificação.

Vale notar que a abordagem baseada em regras não apenas falhou em melhorar a legibilidade, como a reduziu ligeiramente ($-1,6$). Este fenômeno ocorre porque as regras de divisão de sentenças e conversão para voz ativa, embora conceitualmente corretas, introduzem pausas e construções que podem tornar o texto mecanicamente mais pesado quando aplicadas sem consideração do contexto semântico. Por exemplo, a regra de divisão de sentenças com mais de 40 palavras frequentemente fragmenta indevidamente períodos compostos coordenados, criando sentenças curtas mas desconectadas, que exigem maior esforço cognitivo do leitor para reconstruir a relação entre as ideias. Este resultado corrobora as observações de Shardlow [9] sobre a importância do contexto na simplificação lexical e sintática.

5.2 O paradoxo das métricas de similaridade

Um dos achados mais relevantes deste trabalho é a demonstração de que métricas de similaridade textual (ROUGE, BERTScore) não se correlacionam diretamente com legibilidade. A técnica baseada em regras obteve ROUGE-1 de 0,992 e BERTScore F1 estimado de 0,994 — valores próximos do máximo — mas o índice Flesch revelou que a legibilidade efetivamente piorou. Isto ocorre porque métricas baseadas em sobreposição e similaridade semântica premiam a preservação lexical, enquanto a simplificação ideal frequentemente requer paráfrase substancial. Para avaliação de simplificação textual, recomendamos o uso combinado de métricas de legibilidade (Flesch, Gunning-Fog), similaridade (ROUGE) e preservação semântica (BERTScore).

Este paradoxo tem implicações metodológicas importantes. Estudos que utilizam apenas ROUGE ou BERTScore para avaliar simplificação podem superestimar a qualidade de abordagens conservadoras que alteram pouco o texto original. Recomendamos que futuros trabalhos na área adotem, no mínimo, uma métrica de cada dimensão (legibilidade, similaridade lexical e preservação semântica) para capturar os diferentes aspectos da qualidade da simplificação. A recente literatura internacional em simplificação jurídica [20, 15] tem caminhado nessa direção, incorporando métricas de legibilidade específicas do idioma alvo.

5.3 Chain-of-Thought: quando o raciocínio atrapalha

O fraco desempenho do CoT (pior em todas as métricas) desafia a premissa de que raciocínio explícito melhora a geração. No contexto jurídico, as etapas genéricas do CoT podem levar o modelo a ‘‘explicar a decisão’’ em vez de ‘‘reescrever o texto’’. O few-shot, ao fornecer exemplos concretos, mostra-se mais eficaz porque o modelo aprende por analogia o formato esperado da saída. Este resultado sugere que, para tarefas de transformação texto-a-texto em domínios especializados, a demonstração (few-shot) é mais eficaz que a instrução procedural (CoT).

Esta observação está alinhada com a literatura, que reporta que o CoT é mais eficaz em tarefas de raciocínio lógico e matemático do que em tarefas de transformação textual. No entanto, contrasta com os resultados de Wei et al. [28], que demonstraram benefícios do CoT em tarefas de compreensão e geração. A diferença pode ser atribuída à natureza da tarefa: simplificação textual requer reescrita integral, não a adição de raciocínio intermediário. O CoT genérico foi projetado para tarefas analíticas — solução de problemas matemáticos, diagnósticos, raciocínio lógico — em que a decomposição em passos discretos auxilia o modelo a construir a resposta incrementalmente. Na simplificação textual, porém, a tarefa é essencialmente geracional: produzir uma paráfrase acessível de um texto-fonte. Ao ser instruído a ‘‘seguir passos de raciocínio’’, o modelo desloca seu foco da reescrita para a *análise* do texto, gerando um meta-texto explicativo que substitui, em vez de simplificar, o documento original.

Uma hipótese para investigação futura é que um CoT *adaptado ao domínio* — com etapas específicas orientadas à transformação textual em vez de à análise — poderia reverter este resultado negativo. Por exemplo, em vez de passos genéricos como ‘‘identifique as partes envolvidas’’, um CoT adaptado poderia instruir o modelo a (1) *liste os jargões jurídicos e latinismos presentes no texto original*; (2) *para cada termo listado, proponha uma substituição em linguagem simples, preservando o sentido jurídico*; (3) *reescreva o texto integral aplicando as substituições propostas e reorganizando as sentenças para maior clareza*. Este formato de ‘‘rascunho intermediário de simplificações lexicais’’ manteria o modelo focado na transformação do texto em vez de na explicação da decisão, alinhando a estrutura do prompt à natureza geracional da tarefa.

5.4 Trade-off entre públicos-alvo

A análise conjunta das métricas revela que não existe uma técnica universalmente superior. O DeepSeek V4 Flash é a melhor escolha para maximizar legibilidade (ganho de +24,3 pontos), adequado para cidadãos com baixo letramento jurídico que precisam compreender o essencial de uma decisão. O Qwen 2.5 7B (Zero-Shot), com BERTScore F1 de 0,748, é a melhor opção para preservação semântica; o GPT-5.4 (completo) (0,713) e o Nemotron (0,704) são alternativas com maior ganho de legibilidade, sendo mais adequados para contextos onde o equilíbrio entre clareza e precisão é necessário, como na comunicação entre advogados e clientes ou na formação de estudantes de direito. O Big Pickle few-shot oferece um equilíbrio competitivo entre as duas dimensões, com ganho de legibilidade de +24,3 e BERTScore F1 de 0,672, posicionando-se como uma alternativa robusta. O GPT-5.4 Mini ocupa um nicho distinto: embora seu ganho de legibilidade (+16,4) e ROUGE-1 (0,458) sejam inferiores aos do completo, seu tempo de resposta (~2,5 s/doc) o torna a opção mais rápida entre todos os LLMs analisados.

A escolha da técnica ideal depende, portanto, do público-alvo e do contexto de uso. Para um sistema voltado ao cidadão comum em serviços de assistência judiciária, priorizar-se-ia legibilidade (Big Pickle few-shot). Para um sistema de apoio a magistrados e servidores, priorizar-se-ia a preservação semântica (Nemotron). Em cenários reais, um sistema híbrido que combina múltiplos modelos com seleção dinâmica conforme o perfil do usuário e a complexidade do texto poderia oferecer a melhor experiência.

5.5 Escalabilidade: LLMs de grande porte vs. modelos leves

Os resultados também permitem comparar o desempenho de LLMs de grande porte (centenas de bilhões de parâmetros) com o modelo leve Qwen 2.5 7B (7 bilhões de parâmetros), representante de uma classe de modelos mais acessíveis computacional e financeiramente.

O Qwen 2.5 7B, com custo de US$ 0,07/milhão de tokens de entrada e US$ 0,21/milhão de tokens de saída — 10 a 30 vezes menor que os modelos pagos testados — apresentou o menor ganho de legibilidade entre os LLMs (+13,9 no modo Few-Shot e +13,6 no Zero-Shot, contra +24,3 dos líderes). No entanto, obteve a maior preservação semântica (BERTScore mBERT F1 = 0,748 no modo Zero-Shot), superando inclusive modelos muito maiores.

Este resultado sugere que modelos leves como o Qwen 2.5 7B podem ser suficientes para aplicações que priorizam a fidelidade semântica sobre a maximização da legibilidade — por exemplo, ferramentas de apoio a operadores do direito que necessitam de paráfrases fiéis ao texto original. Para cenários que exigem máxima acessibilidade (como sistemas voltados ao cidadão comum com baixo letramento), os modelos de grande porte permanecem superiores, com ganhos de legibilidade quase o dobro.

Em termos de latência, o Qwen 2.5 7B (~9 s/doc) posiciona-se entre os modelos gratuitos via OpenCode Zen (30–40 s/doc) e os modelos otimizados do OpenRouter (2–4 s/doc). Considerando seu custo marginal muito baixo, o Qwen representa uma alternativa viável para processamento em lote de grandes volumes documentais, onde o orçamento é a principal restrição. Adicionalmente, por ser um modelo de pesos abertos, o Qwen 2.5 7B pode ser executado localmente com quantização, eliminando completamente os custos de API e a dependência de conexão com a internet — vantagem significativa para tribunais com restrições de infraestrutura.

5.6 Comparação com Trabalhos Anteriores

Os resultados obtidos são consistentes com a literatura internacional, embora as diferenças metodológicas e linguísticas exijam cautela na comparação direta. [18] analisam o impacto do uso de linguagem simples em documentos jurídicos, documentando ganhos significativos de legibilidade. Embora a comparação direta exija cautela devido às diferenças de idioma e à maior diversidade do nosso corpus, os resultados gerais são consistentes com as tendências reportadas na literatura internacional. [14] demonstraram a eficácia de modelos como GPT-4 em tarefas jurídicas, observando que modelos maiores tendem a apresentar melhor desempenho — achado consistente com nossas observações.

Em contexto internacional, [18] também discutem a simplificação de cláusulas contratuais, reportando ganhos de legibilidade alinhados aos observados em nossos experimentos. Esta diferença pode ser atribuída tanto à maior capacidade dos LLMs atuais quanto à natureza distinta dos textos (cláusulas contratuais vs. decisões judiciais integrais). Silveira et al. [16] desenvolveram recursos de PLN para o domínio jurídico brasileiro, e os ganhos observados em nosso trabalho superam métricas reportadas em tarefas correlatas, sugerindo que a abordagem baseada em LLMs preserva mais similaridade lexical que métodos especializados.

5.7 Implicações Práticas para o Judiciário Brasileiro

Os resultados deste trabalho têm implicações diretas para o movimento de Linguagem Simples no Judiciário brasileiro, que ganhou impulso institucional com a Portaria CNJ n. 351/2023 [2].

Primeiro, a demonstração de que LLMs podem produzir simplificações de qualidade comparável — combinando modelos gratuitos (Big Pickle, Nemotron, DeepSeek, via OpenCode Zen) e pagos (GPT-5.4 Mini, GPT-5.4, Gemini 3.5 Flash, via OpenRouter) — é relevante para tribunais com restrições orçamentárias. O custo total das chamadas de API para geração das 1.600 simplificações foi de US$ 1,94 para o Gemini 3.5 Flash, US$ 1,33 para o GPT-5.4 (completo) e US$ 0,231 para o GPT-5.4 Mini; os modelos do OpenCode Zen não tiveram custo. Em produção, o custo estimado por documento varia entre R$ 0,05 e R$ 0,50, dependendo do modelo e do tamanho do texto, valor marginal quando comparado ao custo de um assistente humano para realizar a mesma tarefa.

Segundo, a identificação de erros específicos (alucinação de valores, generalização excessiva) fornece subsídios para o design de sistemas de simplificação assistida, nos quais o LLM gera uma primeira versão que é revisada por um operador do direito antes da publicação. Este modelo ‘‘humano-no-loop’’ é particularmente adequado para o contexto jurídico, onde a precisão é indispensável.

Terceiro, a variação de desempenho entre temas (consumidor mais favorável, família mais desafiador) sugere que estratégias de simplificação devem ser calibradas por domínio jurídico. Uma abordagem ‘‘one-size-fits-all’’ pode não ser adequada; sistemas reais devem considerar a especialização temática na seleção e configuração dos modelos.

5.8 Considerações Éticas

A simplificação automática de decisões judiciais envolve considerações éticas que não podem ser ignoradas.

O risco mais imediato é a **perda de precisão jurídica** em prol da legibilidade. Uma simplificação que omite uma condição legal importante (como um prazo recursal ou uma exceção legal) pode levar o cidadão a tomar decisões processuais equivocadas. Por este motivo, defendemos que sistemas de simplificação jurídica devem sempre incluir um aviso claro de que o texto simplificado não substitui o texto oficial e que, em caso de dúvida, deve-se consultar um advogado.

A **alucinação de informações** — identificada em duas das 1.600 simplificações geradas — é particularmente grave no contexto jurídico. Diferentemente de outros domínios (como entretenimento ou educação geral), uma informação incorreta em uma simplificação judicial pode ter consequências reais na vida do cidadão, como a perda de um prazo processual ou a tomada de decisão baseada em fatos inexistentes. Sistemas em produção devem implementar mecanismos de verificação factual, idealmente com validação humana.

A **neutralidade e viés** dos modelos também merece atenção. Embora não tenhamos identificado vieses explícitos nas simplificações geradas, a literatura documenta que LLMs podem reproduzir e amplificar vieses dos dados de treinamento [31]. No contexto jurídico, isso poderia se manifestar como simplificações que distorcem argumentos de determinadas partes ou que aplicam tratamento desigual a temas sensíveis, como direito de família ou questões raciais.

Por fim, a **transparência** é fundamental. O cidadão deve ser informado de que o texto foi gerado automaticamente, por qual modelo, e com quais limitações. Defendemos um selo de ‘‘Simplificação Automática’’ nos textos gerados, indicando o modelo e a data de geração, seguindo princípios de explicabilidade e responsabilidade algorítmica.

5.9 Ameaças à Validade

Esta seção discute as ameaças à validade dos resultados segundo a taxonomia de Wohlin et al. [32], organizada em quatro dimensões.

5.9.1 Validade Interna

A validade interna diz respeito à confiança na relação de causalidade entre as técnicas de simplificação e as métricas observadas.

A principal ameaça interna é o **não-determinismo das gerações**: os LLMs foram consultados via API com temperatura 0,3, o que introduz variação entre execuções. Embora a temperatura baixa reduza a variabilidade, réplicas exatas podem produzir diferenças marginais nas métricas. Para mitigar esta ameaça, todas as gerações foram realizadas em bloco único e os resultados foram verificados por consistência entre modos de prompt do mesmo modelo.

A **qualidade da baseline de regras** também representa uma ameaça: o dicionário de substituição lexical contém apenas 30 termos, e as regras de divisão de sentenças e conversão para voz ativa são heuristicamente simples. Uma baseline baseada em regras mais sofisticada poderia alterar a magnitude da diferença observada entre LLMs e regras, embora dificilmente eliminaria a lacuna qualitativa.

Por fim, a **ordem de apresentação dos exemplos** no modo few-shot foi fixa (consumidor, família, previdenciário). Não foi investigado se diferentes ordenações dos exemplos produzem variações na qualidade da simplificação.

5.9.2 Validade Externa

A validade externa refere-se à generalização dos resultados para outros contextos.

O **tamanho e escopo do corpus** são as limitações mais evidentes: 100 decisões de um único tribunal (STF) em três temas jurídicos. Decisões de tribunais estaduais, trabalhistas e eleitorais podem apresentar padrões linguísticos, estilos de redação e níveis de complexidade distintos, limitando a generalização dos achados.

Os **modelos testados** representam uma amostra dos LLMs disponíveis em junho de 2026. Modelos lançados posteriormente, ou não testados (como Claude, Llama 3, Mistral), podem produzir resultados diferentes. Ademais, os modelos via API podem ter sido atualizados entre o período de coleta e a publicação, introduzindo variação não controlada. O estudo concentrou-se exclusivamente em LLMs decoder-only contemporâneos devido à sua ampla adoção prática, não incluindo modelos encoder-decoder tradicionais como mT5 [33], ByT5 [34] ou T5 [35] — arquiteturas consagradas para tarefas de sequência-para-sequência que poderiam oferecer diferentes trade-offs entre legibilidade e preservação semântica.

A **avaliação exclusivamente automática** também limita a validade externa: métricas como Flesch, ROUGE e BERTScore, embora úteis para comparação objetiva, não substituem a avaliação por humanos (juízes, advogados, cidadãos com diferentes níveis de letramento). Estudos futuros devem incluir avaliação humana para validar e complementar as métricas automáticas. Os exemplos qualitativos discutidos na Seção 4.5 são ilustrações autorais elaboradas pelos pesquisadores para demonstrar padrões observados, não resultados de avaliação sistemática com participantes humanos.

5.9.3 Validade de Construto

A validade de construto diz respeito à adequação das métricas para medir o fenômeno de interesse — neste caso, a qualidade da simplificação textual.

O **índice Flesch Adaptado** foi desenvolvido para textos didáticos, não para o discurso jurídico. Sua fórmula, baseada em média de sílabas por palavra e palavras por sentença, pode não capturar adequadamente a complexidade sintática e terminológica de decisões judiciais. Um texto com sentenças curtas mas vocabulário técnico denso pode receber uma pontuação Flesch favorável sem ser genuinamente acessível.

O **BERTScore**, embora correlacionado com avaliação humana em tarefas de geração, utiliza embeddings do mBERT treinados em texto genérico. A similaridade semântica medida pode não refletir a adequação jurídica da simplificação.

O **ROUGE**, por ser baseado em sobreposição de n-gramas, penaliza paráfrases que utilizam vocabulário diferente do original para explicar conceitos jurídicos — precisamente o comportamento desejável em simplificação textual. Como discutido na Seção 4.4, o alto ROUGE das regras (0,992) contrasta com seu ganho negativo de legibilidade, evidenciando este paradoxo.

5.9.4 Validade de Conclusão

A validade de conclusão refere-se à confiança estatística nas inferências realizadas.

O **tamanho amostral** (100 documentos) é adequado para testes não-paramétricos, mas o número de comparações pareadas (120 pares para 16 técnicas) reduz o poder estatístico após a correção de Bonferroni. Algumas diferenças reais entre técnicas podem não ter sido detectadas.

A presença de **outliers extremos** — como o artefato de geração do Qwen 2.5 7B Zero-Shot (repetição infinita em um documento, resultando em ΔFlesch de $-778,6$ antes da correção) — distorce médias e pode afetar testes paramétricos. Para mitigar esta ameaça, adotaram-se testes não-paramétricos (Friedman, Wilcoxon) baseados em *ranks*, robustos a outliers, e reportam-se tanto médias quanto medianas.

O **pressuposto de esfericidade** do teste de Friedman foi verificado indiretamente pela significância dos resultados ($p < 0,001$ para todas as métricas). No entanto, o teste não identifica quais pares diferem, exigindo pós-testes (Nemenyi) que podem ser conservadores.

6 Conclusão

Este artigo apresentou o *LinguagemSimples*, um pipeline computacional para simplificação automática de decisões judiciais brasileiras utilizando modelos de linguagem de grande escala. A comparação sistemática de dezesseis técnicas (sete LLMs com quinze configurações de prompt, mais uma baseline baseada em regras) sobre 100 decisões reais do STF, avaliadas sob 13 métricas complementares, produziu os seguintes achados principais:

1. LLMs superam amplamente a abordagem baseada em regras, demonstrando que a simplificação efetiva exige compreensão semântica e capacidade de reestruturação sintática que algoritmos simbólicos não conseguem reproduzir. Todos os LLMs apresentaram ganho positivo de Flesch, variando de +16,4 (GPT-5.4 Mini) a +24,3 pontos (DeepSeek V4 Flash), contra $-1,6$ das regras.
2. O DeepSeek V4 Flash é a melhor opção para maximizar legibilidade (Flesch 39,7 $\rightarrow$ 64,1), enquanto o Qwen 2.5 7B (Zero-Shot) oferece a melhor preservação semântica (BERTScore F1 = 0,748), seguido pelo GPT-5.4 (completo) (0,713) e Nemotron 3 Ultra (0,704). O Big Pickle few-shot posiciona-se como alternativa robusta (+24,3 Flesch, BERTScore F1 0,672), e o GPT-5.4 Mini oferece o menor tempo de resposta (~2,5 s/doc) com BERTScore F1 de 0,697, combinando eficiência e fidelidade semântica. O Gemini 3.5 Flash completa o conjunto de modelos testados via OpenRouter.
3. O modo Chain-of-Thought mostrou-se contraproducente para este domínio (pior desempenho em todas as métricas), sugerindo que a demonstração por exemplos (few-shot) supera a instrução procedural para tarefas de transformação texto-a-texto.
4. Métricas de similaridade (ROUGE, BERTScore) isoladamente não capturam melhoria de legibilidade, podendo inclusive mascarar a piora na acessibilidade do texto, reforçando a necessidade de avaliação multidimensional com métricas de legibilidade, similaridade e preservação semântica.
5. A análise qualitativa revela padrões de erro específicos dos LLMs (alucinação de valores, generalização excessiva, perda de nuances, invenção de informações processuais) que representam riscos reais para aplicação prática e devem ser endereçados em sistemas de produção com supervisão humana.
6. Os temas jurídicos apresentam desafios distintos: consumidor é o mais favorável à simplificação (ganho de até +28,2 pontos), previdenciário apresenta ganhos moderados (+23,0) e família mostrou-se o mais desafiador (ganho máximo de +22,9), sugerindo que estratégias de simplificação devem ser calibradas por domínio.

O corpus pareado com 100 decisões originais e 1.600 simplificações, bem como o código-fonte completo do pipeline, estão disponíveis publicamente para reprodutibilidade e extensão por outros pesquisadores.

6.1 Trabalhos Futuros

Com base nos resultados e limitações identificados, planejam-se as seguintes direções de pesquisa e desenvolvimento:

Expansão do corpus: incluir decisões de outros tribunais (STJ, TST, TJs estaduais, TRFs), temas adicionais (trabalhista, tributário, penal, eleitoral) e diferentes gêneros textuais (sentenças de primeiro grau, acórdãos, decisões monocráticas), almejando um corpus de milhares de pares que cubra a diversidade linguística do sistema de justiça brasileiro.

Fine-tuning de modelos: utilizar o corpus gerado para fine-tuning supervisionado de modelos abertos (como Llama 3, Mistral e BERT [30]), visando simplificação eficiente, com menor latência e sem dependência de API externa. Modelos fine-tuned especificamente para o domínio jurídico brasileiro podem oferecer melhor equilíbrio entre legibilidade e preservação semântica.

Avaliação humana: realizar estudo controlado com três grupos de participantes — juízes, advogados e cidadãos com diferentes níveis de letramento — para (i) validar as métricas automáticas, (ii) capturar aspectos qualitativos da simplificação que métricas automáticas não mensuram, e (iii) estabelecer padrões-ouro para treinamento de modelos.

Estratégias de prompt adaptativas: desenvolver sistema que seleciona dinamicamente o modo de prompt e o modelo conforme o tema, a complexidade do texto e o perfil do usuário, combinando os pontos fortes de cada estratégia e mitigando as fraquezas identificadas.

Detecção e correção de alucinações: implementar módulo de verificação factual pós-geração que detecte inconsistências entre o texto original e o simplificado, especialmente para valores numéricos, artigos de lei e informações processuais.

Interface web e API pública: implementar aplicação web intuitiva e API REST para uso público da ferramenta, permitindo que cidadãos, defensores públicos e servidores do judiciário simplifiquem decisões judiciais de forma autônoma. A interface deve incluir alertas sobre as limitações do sistema e recomendar consulta a advogado para casos complexos.

Análise de domínio mais ampla: investigar se os padrões observados (superioridade de LLMs, fraqueza do CoT, variação temática) se mantêm em outros gêneros textuais jurídicos, como petições iniciais, contratos, pareceres ministeriais e acórdãos de tribunais superiores.

Agradecimentos

Os autores agradecem à Universidade Federal de São João del-Rei pelo apoio institucional, à equipe do OpenCode Zen pelo acesso aos modelos de linguagem e ao OpenRouter pela infraestrutura de API.

Contribuições dos Autores

J.P.H. Sansão: conceituação, metodologia, implementação, experimentos, análise dos dados, redação do manuscrito original. **M.C.R. Leles**: orientação, supervisão técnica, revisão crítica do manuscrito.

Conflito de interesses

Os autores declaram não haver conflito de interesses.

Disponibilidade de dados

O código-fonte e o corpus estão disponíveis em:
https://www.github.com/jsansao/LinguagemSimples/

Referências

[1] Brasil. Constituição da república federativa do brasil de 1988, 1988.

[2] Conselho Nacional de Justiça. Portaria cnj n. 351/2023: Institui o selo linguagem simples, 2023. Disponível em: https://www.cnj.jus.br.

[3] Conselho Nacional de Justiça. Justiça em números 2022. *Relatório Anual*, 2022. Disponível em: https://www.cnj.jus.br.

[4] Instituto Paulo Montenegro and Ação Educativa. Inaf brasil 2022: Indicador de alfabetismo funcional, 2022. Disponível em: https://ipm.org.br.

[5] IBGE. Pnad contínua: Educação 2021, 2021. Disponível em: https://ibge.gov.br.

[6] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[8] Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. A survey on text simplification. *ACM Computing Surveys*, 55(8):1–36, 2022.

[9] Matthew Shardlow. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70, 2014.

[10] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[11] Hugo Touvron, Louis Martin, Kevin Stone, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[12] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[13] Jean Kaddour, Joshua Harris, Maximilian Mozes, et al. Challenges and applications of large language models. In *arXiv preprint arXiv:2307.10169*, 2023.

[14] Jaromir Savelka and Kevin D. Ashley. The unreasonable effectiveness of large language models in zero-shot semantic annotation of legal texts. *Frontiers in Artificial Intelligence*, 6:1279794, 2023. doi: 10.3389/frai.2023.1279794.

[15] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, 2020. doi: 10.18653/v1/2020.findings-emnlp.261.

[16] Raquel Silveira, Caio Ponte, Vitor Almeida, Vládia Pinheiro, and Vasco Furtado. Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In *Intelligent Systems (BRACIS 2023)*, pages 268–282. Springer, 2023.

[17] Plain Language Action and Information Network. Federal plain language guidelines, 2023. Disponível em: https://www.plainlanguage.gov.

[18] Laura Manor and Junyi Jessy Li. Plain english summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop*, pages 1–11, 2019. doi: 10.18653/v1/W19-2201.

[19] Pedro H. L. Araujo, Teófilo E. de Campos, Renato R. R. Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. Lener-br: A dataset for named entity recognition in brazilian legal text. In *Computational Processing of the Portuguese Language (PROPOR 2018)*, pages 313–323. Springer, 2018.

[20] Antonio Flavio Paula and Celso Camilo-Junior. Evaluating the simplification of Brazilian legal rulings in LLMs using readability scores as a target. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 117–125, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.tsar-1.12.

[21] Daniel Martin Katz, Michael J. Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *SSRN*, 2023. doi: 10.2139/ssrn.4389233. SSRN 4389233.

[22] Sandra Aluísio and Caroline Gasperin. Porsimples: Simplification of portuguese texts fostering digital inclusion and accessibility. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 62–70, 2010.

[23] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: A bilingual bert model for brazilian portuguese. In *Intelligent Systems (BRACIS 2020)*, pages 65–72. Springer, 2020.

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[25] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[26] Teresa B. F. Martins, Claudete M. Ghiraldelo, M. Graças V. Nunes, and Osvaldo N. Oliveira Junior. Readability formulas applied to textbooks in brazilian portuguese. Technical Report 78, ICMSC-USP, 1996. Notas do ICMSC.

[27] Thomas Wolf, Lysandre Debut, Victor Sanh, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[28] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[29] Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

[31] Noam Chomsky, Ian Roberts, and Jeffrey Watumull. The false promise of chatgpt. *The New York Times*, 2023.

[32] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, Björn Regnell, and Anders Wesslén. *Experimentation in software engineering*. Springer, 2nd edition, 2012.

[33] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *NAACL*, 2021.

[34] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Bhargav, and Mihir Kale. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.