Lispector: Fine-tuning de Modelos de Linguagem para Revisão Gramatical e Ortográfica em Português Brasileiro

Andresa Medeiros[1], Felipe Iszlaji[1], Claudia Sarmiento-Moreno[1], Camila Muniz[1], Larissa Ponciano[1], Larissa Dejigov[1], Ronald Monteiro[1], Pedro Kretikouski[1], Guilherme Chaves[1]

[1]Clarice.ai

andresa.medeiros@clarice.ai, felipe@clarice.ai, clasarmor@gmail.com,
camilaamunizz@gmail.com, larissasponciano@gmail.com, contatolarissadc@gmail.com,
ronaldmonteiro.pro@gmail.com, pedro.junior@clarice.ai, guilherme.chaves@clarice.ai

Abstract

Este trabalho apresenta o Lispector, uma família de modelos de linguagem especializados para revisão gramatical e ortográfica em português brasileiro. Comparamos duas estratégias de inferência para a tarefa de correção gramatical de texto com grandes modelos de linguagem (LLMs): (1) *fine-tuning* supervisionado e (2) *prompting few-shot* em modelos de maior escala. Utilizando um conjunto de dados de 4.500 pares de textos reais de usuários (2.500 registros para treino, 1.000 para avaliação e 1.000 para teste), com referências corrigidas por linguistas, analisamos duas variantes do Lispector baseadas em diferentes tamanhos de parâmetros. A avaliação empregou as métricas BLEU, GLEU, METEOR e ROUGE. Os resultados demonstram que modelos menores submetidos a *fine-tuning* supervisionado superam consistentemente em todas as métricas modelos maiores que operam apenas com *prompting*, com o Lispector *small* alcançando ganhos expressivos em métricas de similaridade textual como GLEU (+12%) e BLEU (+13%). Assim, além do aumento de desempenho, os modelos *fine-tuned* apresentam comportamento mais previsível e conservador, características desejáveis em aplicações industriais de escrita assistida. No quesito latência, o Lispector *small* obteve a menor mediana de tempo de resposta entre todos os modelos e o menor P95 entre os *fine-tuned*; o Lispector *large* também se mostrou competitivo. Esses achados indicam que, para tarefas específicas de revisão textual em português brasileiro, o *fine-tuning* pode oferecer vantagens significativas em desempenho e eficiência computacional.

1 Introdução

A correção gramatical e ortográfica automática é uma tarefa fundamental em processamento de linguagem natural (PLN), com aplicações diretas em editores de texto e plataformas educacionais. Em contextos reais, sistemas de revisão textual devem identificar e corrigir erros de modo previsível, conservador e alinhado às normas linguísticas, evitando reformulações desnecessárias que comprometam a experiência do usuário. Trabalhos pioneiros em correção gramatical para português brasileiro utilizaram abordagens baseadas em regras e métodos híbridos, como ReGra (Martins et al., 1998), CoGrOO (Kinoshita et al., 2006) e LanguageTool (Naber, 2003). Com o advento dos grandes modelos de linguagem (LLMs), adotaram-se duas abordagens principais para correção textual: 1) *prompting few-shot* com modelos generalistas como GPT (Coyne et al., 2023; Fang et al., 2023), que não requer treinamento adicional, sendo atraente do ponto de vista industrial por sua facilidade de integração e baixo custo inicial; e (2) *fine-tuning* supervisionado, que adapta pesos de modelos pré-treinados para correção gramatical (Bryant et al., 2023; Rothe et al., 2022), que, embora exija maior investimento em dados e treinamento, permite maior especialização. Embora modelos maiores via *prompting* sejam frequentemente considerados superiores, com viabilidade para português brasileiro (Penteado and Perez, 2023), há evidências limitadas sobre a eficácia comparativa entre *prompting* e *fine-tuning* para correção textual nessa língua. Permanece em aberto até que ponto modelos menores com *fine-tuning* supervisionado podem competir com modelos maiores usando apenas *prompting*. Essa lacuna é relevante para aplicações industriais, com restrições de custo computacional e latência.

2 Metodologia

2.1 Dados

Os dados foram obtidos de textos reais submetidos por usuários na plataforma de edição de texto Clarice.ai. Cada instância consiste em um par formado por texto original e versão corrigida, pro-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 25–29

duzida por linguistas para eliminar erros de ortografia, gramática e pontuação seguindo as normas do português brasileiro e preservando o conteúdo semântico original. O conjunto de dados contém 4.500 registros, divididos em 2.500 para treinamento, 1.000 para avaliação e 1.000 para teste, sem sobreposição entre conjuntos.

2.2 Modelos

Os experimentos compararam dois grupos principais de modelos: (i) a família Lispector, submetida a *fine-tuning* supervisionado para revisão gramatical e ortográfica em português brasileiro, e (ii) modelos de grande escala utilizados como *baselines*, exclusivamente via *prompting few-shot*, sem ajuste adicional de pesos. Importa notar que os modelos, dados e código não estão disponíveis publicamente por envolverem informações proprietárias e confidenciais vinculadas à Clarice.ai. A família Lispector inclui as variantes Lispector *large* (baseada no GPT-4.1), de maior porte, e Lispector *small* (baseada no GPT-4.1 nano). A segunda investiga se um modelo significativamente menor com *fine-tuning* supervisionado poderia alcançar desempenho comparável ou superior ao de modelos maiores que operam apenas via *prompting*. Essa comparação é relevante no contexto industrial, em que restrições de custo, latência e escalabilidade favorecem modelos mais compactos. Como *baselines*, utilizamos os modelos GPT-5, GPT-4.1, GPT-5 nano e GPT-4.1 nano, todos sem *fine-tuning* e avaliados somente via *prompting few-shot*, sendo que os mesmos prompts são usados para todos os modelos de *baseline*. Por razões de propriedade intelectual, os *prompts* específicos utilizados não podem ser divulgados. Em linhas gerais, todos os modelos *baseline* receberam *prompts* estruturados com instruções explícitas para correção gramatical e ortográfica em português brasileiro, incluindo exemplos representativos e orientações para preservar o conteúdo semântico original e evitar reformulações desnecessárias. Todos os modelos foram avaliados sobre o mesmo conjunto de dados de teste, permitindo comparação direta entre *fine-tuning* supervisionado e uso direto de modelos generalistas, e análise do *trade-off* entre capacidade do modelo, especialização para a tarefa e viabilidade industrial.

2.3 Configuração de treinamento

A Tabela 1 apresenta os hiperparâmetros do *fine-tuning* das duas variantes do Lispector. Buscou-se constância de parâmetros entre os modelos, de forma a permitir uma comparação controlada.

2.4 Métricas de avaliação

Os modelos foram avaliados por métricas automáticas escolhidas para contemplar a literatura consolidada de correção gramatical (GEC) e atender a requisitos de aplicações comerciais de edição de texto, como previsibilidade e consistência, evitando reformulações desnecessárias. Utilizamos BLEU (Papineni et al., 2002) e GLEU (Napoles et al., 2015) para medir a precisão de n-gramas, comparando as saídas dos modelos e os textos de referência. O GLEU é particularmente relevante por penalizar alterações desnecessárias, considerando, em simultâneo, a correção de erros e a preservação de trechos corretos. METEOR (Banerjee and Lavie, 2005) foi incluído por considerar correspondências parciais, enquanto ROUGE-1 e ROUGE-2 medem a sobreposição de unigramas e bigramas, respectivamente, permitindo uma análise complementar da preservação estrutural. Adicionalmente, avaliamos a latência de inferência de cada modelo, reportando média, mediana, P95 e valores mínimo e máximo de tempo de resposta em milissegundos. Essas métricas complementam a avaliação de qualidade textual com uma perspectiva de viabilidade operacional, relevante para aplicações industriais com restrições de tempo de resposta. Em conjunto, essas métricas equilibram rigor formal e tolerância a variações linguísticas naturais, importante no português brasileiro, caracterizado por múltiplas formas corretas de realização textual.

3 Resultados

3.1 Comparação geral

A Tabela 2 apresenta os resultados comparativos entre os modelos *baseline* (*prompting few-shot*) e os com *fine-tuning* supervisionado. Observa-se que os modelos da família Lispector superaram de modo consistente todos os *baselines*, indicando maior proximidade lexical e estrutural às correções de referência. Em BLEU, os Lispectors alcançaram 83% contra 70-77% dos GPTs; em GLEU, 89% contra 77-81%. Os modelos Lispector também obtiveram valores superiores em METEOR e ROUGE. Em particular, o desempenho do Lispector *small* é comparável ao do Lispector *large* em quase todas as métricas, apesar do menor número de parâmetros.

A Tabela 3 apresenta os resultados de latência. O

Parâmetro	**Lispector *large***	**Lispector *small***
Modelo base	GPT-4.1	GPT-4.1 nano
Épocas	3	3
Batch size	6	6
Learning rate mult.	0.1	0.1
Tokens treinados	—	~792.000

Table 1: Hiperparâmetros de treinamento.

Modelo	**BLEU**	**GLEU**	**METEOR**	**R-1**	**R-2**
Fine-tuned (zero-shot)					
Lispector *large*	83%	89%	97%	97%	90%
Lispector *small*	83%	89%	97%	97%	89%
Prompting (few shot)					
GPT-5	71%	78%	95%	94%	83%
GPT-4.1	70%	77%	96%	95%	81%
GPT-5 nano	74%	79%	93%	94%	85%
GPT-4.1 nano	77%	81%	94%	94%	87%

Table 2: Resultados comparativos entre modelos fine-tuned e prompting.

Lispector *small* (ft-lispector-small) obteve a menor mediana de tempo de resposta entre todos os modelos (1.378 ms) e o menor P95 entre os modelos *fine-tuned* (2.601 ms), superando inclusive modelos sem *fine-tuning* de porte equivalente. O Lispector *large* (ft-lispector-large) também se mostrou competitivo (mediana de 2.040 ms). Em contraste, os modelos da família GPT-5 apresentaram latências significativamente maiores, com medianas entre aproximadamente 20 e 41 segundos e P95 entre 33 e 162 segundos — valores incompatíveis com uso em produção em editores de texto.

3.2 Análise

A comparação entre *fine-tuning* supervisionado e *prompting few-shot* revela diferenças claras no comportamento dos modelos. Os modelos Lispector, com *fine-tuning* supervisionado, superaram consistentemente todos os modelos operados via *prompting*, com diferenças mais acentuadas em métricas de precisão lexical (BLEU: +13%, GLEU: +12%). Isso sugere que os modelos *fine-tuned* preservam melhor a forma das correções de referência, o que é desejável em aplicações de revisão textual, nas quais alterações desnecessárias compromem a confiança do usuário. Como achado relevante, o Lispector *small*, baseado em um menor tamanho de parâmetros (GPT-4.1 nano), alcançou desempenho equivalente ao Lispector *large* (baseado no GPT-4.1 completo) e superou de forma significativa modelos maiores que operam via *prompting*, em métricas de qualidade e de latência. Em termos de estabilidade, o Lispector *small* apresentou P95 de 2.601 ms, indicando comportamento consistente mesmo nos piores casos — contraste direto com os modelos GPT-5, cujo P95 chega a 162 segundos. Esse conjunto de resultados indica que, para tarefas específicas de revisão textual, a adaptação supervisionada pode ser mais eficaz que o uso direto de modelos generalistas de grande porte, tanto em qualidade quanto em viabilidade operacional.

4 Discussão

4.1 Implicações para aplicações industriais

Os resultados obtidos demonstram que a especialização via *fine-tuning* é um fator relevante para a viabilidade comercial da família Lispector. A família Lispector demonstra maior previsibilidade e alinhamento às correções de referência que modelos generalistas via *prompting*, características desejáveis em sistemas de escrita assistida. O Lispector *small*, com número compacto de parâmetros, alcança desempenho equivalente ao do Lispector *large* e supera, com consistência, modelos maiores sem supervisão — vantagem que se estende também à latência, fator crítico para sistemas de escrita assistida em tempo real. Em termos de latên-

Modelo	Média (ms)	Mediana (ms)	P95 (ms)	Min (ms)	Max (ms)
Fine-tuned (zero-shot)					
ft-lispector-large	3.756	2.040	6.454	1.262	63.676
ft-lispector-small	1.567	1.378	2.601	880	5.435
Prompting (few shot)					
gpt-4.1	2.741	2.556	5.023	1.676	6.789
gpt-4.1-mini	3.277	3.064	5.410	1.853	13.195
gpt-4.1-nano	2.669	2.487	4.033	1.834	4.715
gpt-5	55.015	41.392	162.940	19.265	181.728
gpt-5-mini	21.399	20.792	33.305	8.311	42.597
gpt-5-nano	32.634	31.792	48.946	13.893	76.428

Table 3: Resultados comparativos de latência e tempo de resposta em milissegundos.

cia, em teste com conjunto de dados cujos registros contabilizaram média de 130 tokens, há um contraste severo de latência entre as abordagens, conforme apresentado na Tabela 3. Enquanto o modelo GPT-5 (utilizado via *prompting*) apresenta uma mediana de 41.392 ms e média superior a 55 segundos, tornando-o inviável para interfaces síncronas, o modelo Lispector *small* atinge uma mediana de 1.378 ms. Essa redução drástica de latência permite que a correção ocorra de forma quase instantânea à medida que o usuário digita. O comportamento mais conservador observado nos modelos *fine-tuned* demonstra particular relevância para a experiência do usuário. Os resultados obtidos indicam que o *fine-tuning* favorece correções mais estáveis, pontuais e previsíveis, alinhadas às expectativas de uso em editores de texto. Vale ressaltar que, embora o *fine-tuning* apresente vantagens em desempenho e latência, a abordagem envolve custos iniciais maiores que o *prompting*, incluindo coleta e anotação de dados, treinamento e manutenção do modelo. Para cenários com restrições orçamentárias ou baixo volume de dados disponíveis, o *prompting* pode representar uma alternativa mais acessível. Contudo, os resultados sugerem que, em contextos industriais com demandas de escala e latência, o investimento em *fine-tuning* tende a se justificar. O Lispector está atualmente em operação na plataforma comercial de edição de texto Clarice.ai, processando requisições de usuários reais em escala, o que reforça a viabilidade prática da abordagem proposta.

4.2 Limitações

O conjunto de dados limita-se ao português brasileiro, restringindo a generalização para outras variantes. Além disso, embora os dados englobem uma diversidade natural de gêneros textuais, esta não foi explorada como variável experimental. Outra limitação refere-se à ausência de comparações diretas com ferramentas baseadas em regras, como LanguageTool. A inclusão destes sistemas poderia ampliar a visão de como o Lispector se posiciona no ecossistema de ferramentas de revisão textual. Ademais, os modelos *baseline* foram avaliados com *prompts* padronizados, sem otimização exaustiva; é possível, então, que uma engenharia de *prompt* mais refinada reduzisse parte da diferença observada. A equivalência de desempenho entre Lispector *small* e Lispector *large* indica que modelos compactos são suficientes para a tarefa, embora não permita concluir se *prompts* mais elaborados alcançariam desempenho similar. Ambos os pontos constituem direções relevantes para trabalhos futuros. Por fim, o uso exclusivo de métricas automáticas, embora adequado para comparação em larga escala, não captura com detalhes a aceitabilidade das correções do ponto de vista do usuário final. Reconhecemos que uma análise qualitativa em uma amostra de exemplos poderia enriquecer a interpretação dos resultados, especialmente para identificar casos em que correções válidas mas distintas da referência são penalizadas pelas métricas. Contudo, por se tratar de correção gramatical e ortográfica — e não de revisão estilística —, o espaço de respostas aceitáveis tende a ser mais restrito, o que atenua parcialmente essa limitação.

4.3 Trabalhos futuros

Pretendemos expandir o conjunto de dados com anotações sobre gêneros textuais e contextos comu-

nicativos, como textos acadêmicos, jornalísticos e jurídicos, permitindo investigar estratégias de adaptação mais específicas ao domínio. Outra potencial direção inclui avaliações humanas sistemáticas focadas em critérios de aceitabilidade, fluidez e utilidade percebida pelo usuário. Ademais, estudos em ambientes de produção, incluindo testes A/B com usuários reais, podem fornecer evidências sobre o impacto do *fine-tuning* supervisionado na experiência do usuário.

5 Conclusão

Este trabalho apresentou o Lispector, uma família de modelos de linguagem especializados para revisão gramatical e ortográfica em português brasileiro, desenvolvidos a partir de *fine-tuning* supervisionado em textos reais de usuários. Em uma avaliação comparativa, contrastamos essa abordagem com o uso direto de modelos generalistas de grande escala via *prompting few-shot*, prática amplamente adotada em aplicações industriais. Os resultados demonstram de forma consistente que o *fine-tuning* supervisionado permite ganhos expressivos de desempenho, mesmo em modelos menores. O Lispector *small* superou modelos significativamente maiores utilizados via *prompting*, com ganhos de até 13% em BLEU e 12% em GLEU. Ademais, o Lispector *small* e o Lispector *large* de destacaram no quesito latência. Isso sugere que, para tarefas de correção gramatical e ortográfica, a adaptação via *fine-tuning* oferece vantagens em desempenho e eficiência computacional. O uso atual do Lispector em uma interface de edição de texto reforça a aplicabilidade da abordagem em cenários reais, evidenciando que modelos compactos *fine-tuned* podem atender de forma eficaz às demandas de sistemas industriais de escrita assistida.

6 Agradecimentos

Este trabalho foi financiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) por meio de bolsas de Treinamento Técnico. Agradecemos à Clarice.ai e a seus usuários, que contribuíram com os dados utilizados.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, page 1–59.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. *Preprint*, arXiv:2303.14342.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *Preprint*, arXiv:2304.01746.

Jorge Kinoshita, Laís do Nascimento Salvador, and Carlos Eduardo Dantas de Menezes. 2006. CoGrOO: a Brazilian-Portuguese grammar checker based on the CETENFOLHA corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Ronaldo Teixeira Martins, Ricardo Hasegawa, Maria Das Graças VolpeNunes, Gisele Montilha, and Osvaldo Novais De Oliveira. 1998. Linguistic issues in the development of regra: A grammar checker for brazilian portuguese. *Nat. Lang. Eng.*, 4(4):287–307.

Daniel Naber. 2003. Languagetool.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 588–593, Beijing, China. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maria Carolina Penteado and Fábio Perez. 2023. Evaluating gpt-3.5 and gpt-4 on grammatical error correction for brazilian portuguese. *Preprint*, arXiv:2306.15788.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2022. A simple recipe for multilingual grammatical error correction. *Preprint*, arXiv:2106.03830.