

SLMs como Avaliadores Automatizados: Avaliando Competências de Escrita Narrativa com Modelos Adaptados por LoRA

**Hyan Batista¹, Lenon Anthony¹, Andreza Falcão¹,
Gabriel Alves¹, Maria da Conceicao Moraes Batista¹,
André C. A. Nascimento¹, Rafael Ferreira Mello¹**

¹Universidade Federal Rural de Pernambuco (UFRPE)
Recife – PE – Brasil

{hyan.batista, lenon.anthony, andreza.mcfalcao}@ufrpe.br
{gabriel.alves, maria.cmbatista}@ufrpe.br
{andre.camara, rafael.mello}@ufrpe.br

Abstract. *Assessing narrative essays in elementary education is pedagogically relevant, but costly, time-consuming, and subject to evaluator variability. As LLMs have been studied for this task but are costly to run, this paper investigates whether small language models (SLMs), adapted with LoRA and executed locally with 4-bit quantization, can support automatic scoring of Brazilian Portuguese narrative essays. Six models from the SmolLM, Gemma-3, and Qwen2.5 families, ranging from 135M to 1B parameters, were evaluated on a corpus annotated in four writing competencies: formal register, thematic coherence, narrative rhetorical structure, and cohesion. We compare LoRA-adapted models with two zero-shot baselines: the original task prompt and a strict JSON-oriented prompt. Results show that zero-shot prompting is unreliable for end-to-end scoring: the original prompt produced no valid structured outputs, while the strict prompt improved validity for some models but yielded poor scoring performance. In contrast, all LoRA-adapted SLMs generated valid outputs for the full test set. The best model, SmolLM-135M-4bit, achieved 0.617 accuracy, 90.5% off-by-one accuracy, and 0.480 MAE. These findings suggest that quantized SLMs adapted with LoRA are a viable alternative for automatic essay scoring in computationally constrained educational settings.*

Resumo. *A avaliação de redações narrativas no ensino fundamental é pedagogicamente relevante, mas custosa, demorada e sujeita à variabilidade entre avaliadores. Como LLMs já foram estudados nesta tarefa, mas têm inferência custosa, este trabalho investiga se pequenos modelos de linguagem (SLMs), adaptados com LoRA e executados localmente com quantização de 4 bits, podem apoiar a correção automática de redações narrativas em português brasileiro. Foram avaliados seis modelos das famílias SmolLM, Gemma-3 e Qwen2.5, de 135M a 1B parâmetros, em um corpus anotado em quatro competências: registro formal, coerência temática, estrutura retórica narrativa e coesão. Comparamos modelos adaptados por LoRA com duas baselines zero-shot: o prompt original da tarefa e um prompt estrito orientado à geração de JSON válido. Os resultados mostram que o prompting zero-shot é pouco confiável para avaliação fim-a-fim: o prompt original não produziu saídas estruturadas válidas, enquanto o prompt estrito melhorou a validade em alguns modelos, mas manteve*

baixo desempenho de pontuação. Em contraste, todos os SLMs adaptados por LoRA geraram saídas válidas para todo o conjunto de teste. O melhor modelo, SmolLM-135M-4bit, obteve acurácia de 0,617, acerto off-by-one de 90,5% e MAE de 0,480. Os achados sugerem que SLMs quantizados e adaptados por LoRA são uma alternativa viável para avaliação automática de redações em contextos educacionais com restrição computacional.

1. Introdução

Escrever é uma forma de organizar experiências, sustentar ideias, narrar acontecimentos e participar de práticas sociais mediadas pela linguagem. No ensino fundamental, a produção de textos narrativos desempenha um papel formativo importante, pois torna visíveis diferentes aspectos do desenvolvimento da escrita, como a relação com o tema proposto, a construção de eventos, a progressão textual, o uso de conectores e o registro linguístico escolhido pelo estudante [Barrera and Santos 2016]. Esses aspectos aparecem recorrentemente na literatura sobre o ensino de escrita e em práticas baseadas em evidências [Graham and Harris 2019]. Ao mesmo tempo, avaliar redações com cuidado é uma tarefa trabalhosa para os professores, especialmente quando envolve turmas numerosas, múltiplos critérios e a necessidade de acompanhamento contínuo dos estudantes [Ferreira et al. 2024].

Sistemas de avaliação automática de redações (*Automated Essay Scoring*, AES) têm sido investigados como forma de apoiar processos avaliativos em larga escala [Shermis and Burstein 2013]. Atualmente, modelos de linguagem de grande porte (LLMs) ampliaram o repertório desses sistemas, pois permitem avaliações orientadas por instruções, adaptação textual a rubricas e geração de feedback em linguagem natural [Zhao et al. 2026, Kasneci et al. 2023, Liew and Tan 2024]. No entanto, quando a avaliação envolve dados educacionais de crianças e adolescentes, o uso de modelos proprietários na nuvem levanta preocupações práticas e éticas. Entre elas estão o custo recorrente, a dependência de conectividade, a governança dos dados e a privacidade [Slade and Prinsloo 2013, Kasneci et al. 2023]. Tais questões se acentuam nas redes públicas e em escolas com infraestrutura computacional limitada.

Nesse cenário, pequenos modelos de linguagem (*Small Language Models*, SLMs) mostram-se adequados por demandarem menos recursos computacionais e permitirem aplicações mais acessíveis [Nguyen et al. 2025]. Revisões recentes da área indicam que SLMs têm recebido atenção crescente por sua eficiência, menor custo de inferência e adequação a cenários *on-device* [Lu et al. 2024]. Quando combinados com quantização e adaptação eficiente de parâmetros, como LoRA [Hu et al. 2022] e QLoRA [Dettmers et al. 2023], esses modelos podem reduzir a dependência de serviços externos e favorecer soluções de *learning analytics* mais controláveis no contexto escolar [Siemens 2013]. Contudo, a eficácia desses modelos na avaliação de redações, especialmente em idiomas pouco representados, como o português brasileiro, ainda não foi amplamente explorada na literatura.

Este trabalho investiga o ajuste fino de SLMs com adaptadores LoRA para a correção automática de redações narrativas em português. Avaliamos seis modelos das famílias SmolLM, Gemma-3 e Qwen2.5, com tamanhos entre 135 milhões e 1 bilhão de parâmetros, em uma tarefa de pontuação multidimensional. Cada redação recebe notas

de 1 a 5 em quatro competências: registro formal, coerência temática, estrutura retórica-narrativa e coesão. Além dos modelos ajustados, o estudo inclui duas linhas de base zero-shot, o que permite separar a capacidade de produzir uma saída estruturada da capacidade de aplicar a rubrica de avaliação. Para conduzir a investigação de forma estruturada, o experimento foi organizado em torno de três perguntas de pesquisa.

QUESTÃO DE PESQUISA 1 (QP1):

LMs sem ajuste fino conseguem produzir saídas estruturadas válidas?

QUESTÃO DE PESQUISA 2 (QP2):

O ajuste fino com LoRA permite atribuir notas próximas às avaliações humanas?

QUESTÃO DE PESQUISA 3 (QP3):

Quais competências de escrita narrativa são mais difíceis para os modelos?

2. Trabalhos Relacionados

A avaliação automática de redações tem sido amplamente explorada na literatura. O trabalho seminal de Page [Page 1966] já antecipava a possibilidade de correção computacional de textos escritos. Desde então, os sistemas de AES evoluíram de métodos baseados em atributos linguísticos e em modelos estatísticos para abordagens neurais e, mais recentemente, para modelos de linguagem de grande porte. Revisões e compilações da área indicam que AES pode apoiar professores ao reduzir o tempo de correção e viabilizar feedback mais frequente, embora desafios de validade, interpretabilidade e alinhamento com critérios pedagógicos permaneçam centrais [Shermis and Burstein 2013, Wang and Wang 2025].

No contexto brasileiro, estudos têm investigado a correção automática de redações em português em diferentes gêneros e níveis educacionais. O corpus de redações narrativas usado neste trabalho foi descrito como um conjunto de dados multicompetência para AES em português brasileiro [Oliveira et al. 2025] e disponibilizado publicamente no Kaggle [Silva Filho and collaborators 2024]. A competição PROPOR'24 avaliou abordagens para pontuação automática dessas narrativas e reportou níveis de concordância de justos a moderados, destacando a dificuldade da tarefa [Mello et al. 2024]. Trabalhos relacionados também investigaram competências específicas, como registro formal em narrativas de estudantes [Silva Filho et al. 2023].

Com o avanço dos LLMs, pesquisas recentes passaram a investigar seu uso em AES e na avaliação educacional. Liew e Tan [Liew and Tan 2024] avaliaram GPT-3.5 em redações do IELTS e destacaram o papel do *prompt engineering*. Seßler et al. [Seßler et al. 2025] compararam LLMs com avaliações docentes em uma pontuação multidimensional, observando bons resultados em alguns critérios, mas também limitações em aspectos complexos, como a coerência e a lógica textual. Hou et al. [Hou et al. 2025] investigaram o enriquecimento de prompts com características linguísticas para melhorar o desempenho do AES em LLMs. Um estudo recente avaliou LLMs proprietários, incluindo Gemini, GPT-4o, Claude e Mistral, na correção de redações narrativas em português brasileiro do ensino fundamental [Lobo et al. 2025]. Esses estudos reforçam que prompting influencia o desempenho, mas não elimina a necessidade de avaliação sistemática.

Diferentemente dos trabalhos anteriores, o presente estudo enfoca SLMs, execução local, quantização de 4 bits e adaptação por meio de LoRA. Assim, desloca a análise de modelos proprietários na nuvem para modelos leves, com potencial de execução em hardware de borda. Até onde verificamos, a literatura recente sobre SLMs e modelos sub-bilhão tem enfatizado o seu uso em cenários de inferência local e dispositivos com restrição e latência [Liu et al. 2024, Allal et al. 2025]. Essa direção também aparece nas famílias de modelos utilizadas neste estudo, como SmolLM, Gemma3, Qwen 2.5 [Hugging Face TB 2024, Gemma Team 2025, Qwen Team 2024], que disponibilizam variantes compactadas entre 135M e 1B de parâmetros. Combinadas com quantização e adaptação eficientes por LoRA, essas arquiteturas vão permitir investigar uma alternativa à dependência de modelos proprietários em nuvem.

3. Método

A Figura 1 apresenta uma visão geral do pipeline experimental. O fluxo parte dos dados anotados, passa pela preparação dos prompts e se divide em dois ramos: avaliação zero-shot, sem atualização de pesos, e ajuste fino supervisionado com LoRA. As duas saídas são avaliadas pelo mesmo validador de JSON e pelas mesmas métricas.

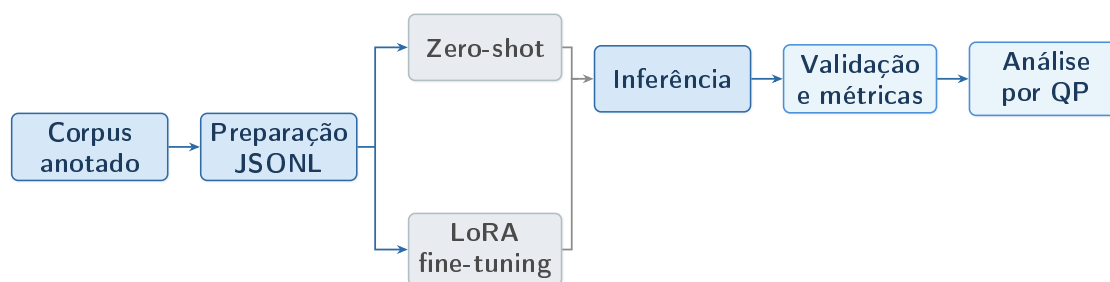


Figura 1. Pipeline experimental utilizado para avaliar modelos zero-shot e ajustados com LoRA.

3.1. Banco de dados

O estudo utiliza um corpus de redações narrativas em português brasileiro produzidas por estudantes do ensino fundamental. O conjunto foi disponibilizado no Kaggle como *Portuguese Narrative Essays* e descrito em uma publicação de dados sobre o dataset de AES multicompetência [Oliveira et al. 2025, Silva Filho and collaborators 2024]. Cada redação está associada a um tema motivador e a notas humanas em quatro competências: registro formal, coerência temática, estrutura retórica narrativa e coesão. As notas variam de 1 a 5, em que 1 representa menor desempenho e 5 maior desempenho na competência.

A Tabela 1 resume a interpretação operacional das quatro competências. A rubrica segue a organização adotada no corpus e em trabalhos anteriores sobre o mesmo domínio [Oliveira et al. 2025, Lobo et al. 2025]. O objetivo do modelo não é fornecer uma explicação pedagógica completa, mas sim atribuir uma pontuação estruturada a cada dimensão.

Tabela 1. Competências avaliadas na tarefa.

Competência	Interpretação operacional
Registro formal	Adequação ao registro escrito esperado para a situação escolar, incluindo ortografia, pontuação, concordância, segmentação, escolha lexical e construções morfosintáticas.
Coerência temática	Relação da narrativa com o tema motivador, manutenção do tópico, progressão das ideias e ausência de rupturas semânticas que comprometam a compreensão global.
Estrutura retórica narrativa	Organização dos elementos narrativos, como situação inicial, personagens, eventos, conflito ou complicação, desenvolvimento temporal e desfecho.
Coesão	Uso de mecanismos que conectam partes do texto, incluindo retomadas referenciais, conectores, continuidade entre sentenças e encadeamento local dos eventos.

O conjunto foi dividido em treino, validação e teste, com 740, 125 e 370 redações, respectivamente [Oliveira et al. 2025, Silva Filho and collaborators 2024]. A Tabela 2 resume a distribuição. O conjunto de teste foi mantido fixo para todas as avaliações, incluindo modelos ajustados e baselines zero-shot.

Tabela 2. Distribuição dos exemplos por partição [Oliveira et al. 2025].

Partição	Número de redações	Uso no experimento
Treino	740	Ajuste fino com LoRA
Validação	125	Monitoramento durante treinamento
Teste	370	Avaliação final

Os dados foram convertidos para o formato JSONL, contendo uma instrução, o tema, a redação e a saída esperada, em um único objeto JSON. A saída-alvo contém exatamente as quatro chaves correspondentes às competências avaliadas: *formal_register*, *thematic_coherence*, *narrative_rhetorical_structure* e *cohesion*.

3.2. Modelos avaliados

Foram avaliados seis SLMs das famílias SmolLM, Gemma-3 e Qwen2.5. A seleção buscou cobrir modelos recentes com entre 135 milhões e 1 bilhão de parâmetros, incluindo variantes quantizadas em 4 bits quando disponíveis. Estes foram escolhidos por serem recentes e pequenos o suficiente para execução local. Nesse processo, foram priorizados modelos abertos publicamente e disponíveis a modo a permitir reprodutibilidade e execução local sem dependência de APIs proprietárias. As famílias diferentes de modelos nos permite comparar arquiteturas diferentes e tokenizadores diversos. A Tabela 3 apresenta os modelos.

Tabela 3. Modelos avaliados.

Identificador no experimento	Família / variante
smol-lm-135m-4bit	SmolLM 135M, 4 bits
smol-lm-135m-it-4bit	SmolLM 135M Instruct, 4 bits
smol-lm-360m-it	SmolLM 360M Instruct
gemma-3-270m-it-4bit	Gemma-3 270M Instruct, 4 bits
gemma-3-1b-it-4bit	Gemma-3 1B Instruct, 4 bits
qwen2.5-0.5b-it-4bit	Qwen2.5 0.5B Instruct, 4 bits

3.3. Ajuste fino com LoRA

O ajuste fino foi realizado com LoRA, um método de adaptação de baixa dimensionalidade para modelos pré-treinados [Hu et al. 2022]. Em vez de atualizar todos os pesos de uma camada linear W , LoRA mantém W congelado e aprende uma atualização de baixo posto $\Delta W = BA$, em que A e B são matrizes treináveis com posto r muito menor que as dimensões originais da camada. Na prática, a ativação passa a usar $Wx + \alpha BAx/r$, em que α controla a escala da atualização. Isso reduz o número de parâmetros treináveis, o consumo de memória e o custo de armazenamento dos adaptadores, preservando o modelo base.

Essa escolha é adequada ao desenho experimental porque a tarefa é especializada, o conjunto de treino contém apenas 740 redações, e o objetivo é manter a inferência local viável. Como citado no parágrafo acima, os modelos base foram congelados e apenas os adaptadores LoRA foram treinados, o que nos permitiu alcançar um método de adaptação eficiente para a tarefa [Hu et al. 2022]. A configuração experimental adotou 100 iterações de treinamento, taxa de aprendizado de 5×10^{-5} , tamanho de lote de 4, comprimento máximo de sequência de 1000 tokens e LoRA com rank 32, escala 32 e dropout de 0,1. As saídas foram limitadas a poucos tokens, pois a tarefa exige apenas um objeto JSON curto com quatro notas.

3.4. Baselines zero-shot

Para avaliar se os modelos seriam capazes de resolver a tarefa sem ajuste fino, foram consideradas duas baselines zero-shot. A primeira, denominada *zero-shot original*, utiliza a instrução original do pipeline de dados. Essa condição testa se o modelo, sem treinamento supervisionado na tarefa, consegue seguir o mesmo protocolo de entrada e saída utilizado no ajuste fino. A segunda, denominada *zero-shot strict*, usa uma instrução reconstruída com restrições explícitas de saída: responder exclusivamente em JSON válido, sem texto adicional, sem Markdown e com as quatro chaves exatas esperadas. A Tabela 4 apresenta as duas instruções.

A baseline original avalia a capacidade de seguir o prompt efetivamente utilizado no pipeline. Uma baseline estrita oferece uma condição mais favorável para modelos sem treinamento, reduzindo ambiguidades quanto ao formato da resposta. Assim, torna-se possível distinguir falhas de formatação de falhas na aplicação da rubrica.

Tabela 4. Prompts usados nas baselines zero-shot.

Condição	Estrutura do prompt
Instrução	Você é um especialista no ensino de redação narrativa para alunos do ensino fundamental. Dada uma determinada entrada (uma redação narrativa), seu papel é avaliá-la nas seguintes competências: registro formal, coerência temática, estrutura retórica narrativa e coesão. Sua nota deve ir de 1 à 5, sendo 1 a menor nota e 5 a maior. A saída deve ser um dicionário com as notas de cada competência, por exemplo: "formal_register": 5, "thematic_coherence": 2, "narrative_rhetorical_structure": 4, "cohesion": 3 Tema: {Tema} {redação}.
Strict	Você é um avaliador de redações narrativas em português brasileiro do ensino fundamental. Avalie a redação abaixo em quatro competências: 1. formal_register 2. thematic_coherence 3. narrative_rhetorical_structure 4. cohesion. Cada competência deve receber uma nota inteira de 1 a 5. Responda exclusivamente com um objeto JSON válido. Não inclua explicações. Não use markdown. Não escreva texto antes ou depois do JSON. Use exatamente estas quatro chaves: "formal_register", "thematic_coherence", "narrative_rhetorical_structure", "cohesion". Formato obrigatório: "formal_register": 1, "thematic_coherence": 1, "narrative_rhetorical_structure": 1, "cohesion": 1 Tema: {Tema} {redação}.

3.5. Avaliação

Cada saída gerada foi considerada válida apenas quando contivesse um objeto JSON válido, ou seja, com todas as quatro chaves esperadas e valores inteiros no intervalo de 1 a 5. Respostas ausentes, malformadas, com chaves ausentes ou valores fora da escala foram consideradas inválidas.

A avaliação foi organizada de modo a responder às perguntas de pesquisa. Para QP1, reportamos as taxas de saída válidas, pois modelos que não seguem o formato esperado não podem ser avaliados como corretores automáticos fim-a-fim. Para QP2 e QP3, calculamos as métricas por modelo e por competência, permitindo comparar o efeito do ajuste do LorA, o comportamento das famílias avaliadas e a dificuldade relativa das quatro dimensões da rubrica.

Para saídas válidas, calculamos acurácia, acurácia com tolerância de um ponto (*off-by-one accuracy*) e erro absoluto médio (MAE). A acurácia mede a proporção de notas exatamente iguais às notas humanas. A métrica *off-by-one* considera correta uma predição cuja diferença absoluta em relação à nota humana é menor ou igual a um ponto, capturando aproximações aceitáveis em uma escala ordinal de 1 a 5. O MAE mede a distância média entre a nota predita e a nota humana, preservando a magnitude do erro. Essas métricas são adequadas para avaliação automática de redações porque combinam correspondência exata, tolerância ordinal e distância média entre avaliador humano e modelo, estratégia também observada em trabalhos anteriores de pontuação automática de escrita [Mello et al. 2024, Lobo et al. 2025, Oliveira et al. 2025]

4. Resultados

Os resultados respondem às três perguntas de pesquisa, começando pela validade formal das saídas e avançando para o desempenho, o tamanho do modelo e a variação entre

competências.

4.1. QP1: Saídas estruturadas sem ajuste fino

A QP1 pergunta se SLMs sem ajuste fino conseguem operar no protocolo de saída da tarefa. A Tabela 5 apresenta a taxa de saídas válidas nas três condições avaliadas: zero-shot com o prompt original, zero-shot com o prompt estrito e modelos ajustados com LoRA.

Tabela 5. Taxa de saídas válidas por modo de avaliação.

Modelo	Zero-shot original	Zero-shot strict	LoRA
gemma-3-1b-it-4bit	0/370	19/370	370/370
gemma-3-270m-it-4bit	0/370	318/370	370/370
qwen2.5-0.5b-it-4bit	0/370	370/370	370/370
smol-lm-135m-4bit	0/370	0/370	370/370
smol-lm-135m-it-4bit	0/370	0/370	370/370
smol-lm-360m-it	0/370	0/370	370/370

Com o prompt original, nenhum modelo zero-shot produziu saídas válidas. A resposta à QP1, portanto, é negativa nessa condição: os modelos base não conseguiram seguir o formato de saída do pipeline de treinamento. Com o prompt estrito, a validade formal melhorou em alguns modelos: Qwen2.5-0.5B gerou 370 saídas válidas, Gemma-3-270M gerou 318 e Gemma-3-1B gerou 19. No entanto, os modelos SmoLLM permaneceram com 0 saídas válidas. Após o ajuste fino com LoRA, todos os modelos geraram saídas válidas em todos os exemplos de teste.

4.2. QP2: Desempenho dos modelos ajustados

A QP2 avalia se, uma vez resolvido o problema de formato de saída, os modelos ajustados por LoRA produzem notas próximas às avaliações humanas. A Tabela 6 resume o desempenho geral dos modelos ajustados. O melhor resultado foi obtido pelo SmoLLM-135M-4bit, com acurácia de 0,617, taxa off-by-one de 90,5% e MAE de 0,480. Em seguida, aparecem SmoLLM-135M-it-4bit, com acurácia de 0,608, off-by-one de 88,5% e MAE de 0,511, e Gemma-3-1B-it-4bit, com acurácia de 0,607, off-by-one de 88,4% e MAE de 0,513.

Tabela 6. Desempenho geral dos modelos ajustados com LoRA.

Modelo	Acurácia	Off-by-one	MAE
smol-lm-135m-4bit	0,617	0,905	0,480
smol-lm-135m-it-4bit	0,608	0,885	0,511
gemma-3-1b-it-4bit	0,607	0,884	0,513
qwen2.5-0.5b-it-4bit	0,595	0,882	0,528
smol-lm-360m-it	0,565	0,879	0,561
gemma-3-270m-it-4bit	0,534	0,855	0,630

A resposta à QP2 é positiva, mas limitada. O ajuste LoRA levou todos os modelos a produzir respostas válidas e a alcançar taxas off-by-one superiores a 85%, indicando que

a maioria das predições ficou a, no máximo, um ponto da nota humana. O menor MAE foi obtido pelo SmolLM-135M-4bit, com 0,480, o que corresponde a um erro médio inferior a meio ponto na escala de 1 a 5. No entanto, a maior acurácia exata foi de 0,617, mostrando que ainda há divergências frequentes entre as notas preditas e as notas humanas. Portanto, os resultados indicam uma aproximação ordinal útil, mas não sustentam o uso autônomo sem supervisão docente.

4.3. QP3: Desempenho por competência

A QP3 trata da dificuldade relativa das quatro dimensões da rubrica. A Tabela 7 apresenta o MAE por competência para os modelos ajustados. A competência de coerência temática foi a mais difícil, com MAE entre 0,819 e 1,170. Por outro lado, a coesão e o registro formal apresentaram valores mais baixos, especialmente no SmolLM-135M-4bit.

Tabela 7. MAE por competência nos modelos ajustados com LoRA.

Modelo	Registro formal	Coerência temática	Estrutura narrativa	Coesão	Geral
smol-lm-135m-4bit	0,362	0,819	0,400	0,341	0,480
smol-lm-135m-it-4bit	0,365	0,897	0,422	0,362	0,511
gemma-3-1b-it-4bit	0,365	0,903	0,422	0,362	0,513
qwen2.5-0.5b-it-4bit	0,365	0,903	0,481	0,362	0,528
smol-lm-360m-it	0,365	0,889	0,630	0,362	0,561
gemma-3-270m-it-4bit	0,465	1,170	0,422	0,462	0,630

O SmolLM-135M-4bit obteve o menor MAE em todas as competências, com destaque para a coesão (0,341) e o registro formal (0,362). A resposta à QP3 é que a dificuldade não se distribui uniformemente: a coerência temática foi a dimensão mais instável, possivelmente por exigir julgamento global sobre a manutenção do tema, a progressão textual e a relação entre a narrativa produzida e a proposta. Essa dificuldade é coerente com estudos de AES que observam maior desempenho em aspectos mais locais ou formais do texto do que em critérios globais de conteúdo e organização [Seßler et al. 2025, Hou et al. 2025].

5. Discussão

As respostas às perguntas de pesquisa indicam uma diferença clara entre seguir o formato da tarefa e pontuar a redação de forma alinhada à rubrica. Em relação à QP1, os modelos zero-shot não foram confiáveis como avaliadores estruturados: o prompt original não produziu JSON válido, e o prompt estrito apenas resolveu parcialmente o problema de formato. Em relação à QP2, o ajuste com LoRA mudou esse comportamento, pois todos os modelos ajustados passaram a gerar respostas válidas em todo o conjunto de teste e alcançaram off-by-one acima de 85%. Essa melhora, porém, não deve ser interpretada como forte concordância com avaliadores humanos: o melhor modelo obteve acurácia de 0,617 e MAE de 0,480.

A QP3 aponta para uma limitação pedagógica, em que a coerência temática foi a competência mais difícil, provavelmente por exigir comparação entre o tema motivador e o desenvolvimento global da narrativa, enquanto o registro formal e a coesão podem depender mais de pistas locais. Além disso, inspeções das predições indicam que alguns modelos tendem a concentrar as respostas nas classes majoritárias. Esse comportamento

pode inflar métricas como acurácia e off-by-one em distribuições desbalanceadas, sem necessariamente indicar uma compreensão profunda da rubrica. Por isso, a leitura conjunta de acurácia, off-by-one e MAE é mais adequada do que a escolha de uma única métrica.

Do ponto de vista educacional, a proposta tem potencial para apoiar professores em atividades de triagem, diagnóstico e priorização de devolutivas, principalmente em contextos em que a correção manual frequente é inviável. Contudo, o uso pedagógico deve preservar a supervisão humana. O sistema deve ser entendido como apoio à avaliação e não como substituto da mediação docente, especialmente porque aspectos de criatividade, intencionalidade e contexto sociolinguístico podem escapar a modelos automáticos.

6. Ameaças à Validade

Há ameaças à validade interna e externa. Primeiro, o conjunto de dados apresenta desbalanceamento entre as classes, o que pode favorecer modelos que predizem as classes majoritárias. Por esse motivo, reportamos acurácia, off-by-one, MAE e análise por competência, mas ainda é necessário investigar estratégias de balanceamento, reamostragem ou calibração ordinal.

Em segundo lugar, os resultados dependem da configuração de prompt, da tokenização e da decodificação. A baseline zero-shot strict melhora a validade formal em alguns modelos, mas não utiliza decodificação gramaticalmente restrita. Técnicas de geração com restrição de JSON poderiam aumentar a taxa de validade, embora não necessariamente melhorassem a qualidade das notas.

Terceiro, os experimentos foram conduzidos com uma única partição de treino, validação e teste. Embora essa partição permita uma comparação controlada entre modelos, validações cruzadas ou múltiplas sementes poderiam fornecer estimativas mais robustas. Quarto, o estudo avalia apenas redações narrativas em português brasileiro de estudantes do ensino fundamental; portanto, os resultados não devem ser generalizados diretamente a outros gêneros textuais, anos escolares ou contextos avaliativos.

Por fim, os modelos ajustados foram avaliados quanto de correspondência com as notas humanas do corpus. O trabalho não avalia o impacto pedagógico em sala de aula, a aceitação por parte dos professores ou os efeitos sobre a aprendizagem dos estudantes. Essas dimensões são essenciais antes de qualquer implantação prática.

7. Conclusão

Este trabalho investigou o uso de pequenos modelos de linguagem adaptados por LoRA para avaliação automática de competências de escrita narrativa em português brasileiro. Foram avaliados seis modelos, com entre 135 milhões e 1 bilhão de parâmetros, com foco na execução local e na quantização de 4 bits. Os modelos ajustados foram comparados a baselines zero-shot com o prompt original e o prompt estrito.

As perguntas de pesquisa foram respondidas da seguinte forma. Para a QP1, os modelos zero-shot não foram suficientes para operar no protocolo de saída: o prompt original não gerou saídas válidas, e o prompt estrito apenas aumentou a validade formal em parte dos modelos. Para a QP2, o ajuste LoRA foi necessário para tornar a avaliação fim-a-fim operacional; todos os modelos ajustados geraram respostas válidas em todo o conjunto de teste. O SmolLM-135M-4bit obteve o melhor desempenho geral, com acurácia

de 0,617, acerto off-by-one de 90,5% e MAE de 0,480. Para a QP3, a coerência temática foi a competência mais difícil, enquanto a coesão, a estrutura narrativa e o registro formal apresentaram resultados mais elevados.

Esses achados sugerem que SLMs quantizados, quando adaptados por LoRA, constituem uma alternativa viável para apoiar a avaliação automática de redações em ambientes com restrições computacionais. Como trabalhos futuros, pretende-se investigar métodos de balanceamento, calibração ordinal, geração estruturada com restrições formais, uso de múltiplas sementes e estudos com professores para avaliar a utilidade pedagógica e a interpretabilidade das notas.

Declaração sobre uso de Inteligência Artificial

Reconhecemos o uso de IA Generativa para auxiliar na elaboração e revisão deste artigo, que foi devidamente revisado, e assumimos total responsabilidade pelo seu conteúdo.

Disponibilidade de Artefatos

Os artefatos experimentais incluem notebooks de preparação de dados, treinamento com LoRA, avaliação dos modelos, baselines zero-shot e agregação de métricas, além dos resultados gerados para cada modelo: Notebooks: <https://anonymous.4open.science/r/nslms-A08D/notebooks>.

Referências

- Allal, L. B., Lozhkov, A., Bakouch, E., Blázquez, G. M., Penedo, G., Tunstall, L., Marafioti, A., Kydlíček, H., Piqueres Lajarín, A., Srivastav, V., Lochner, J., Fahlgren, C., Nguyen, X.-S., Fourrier, C., Burtenshaw, B., Larcher, H., Zhao, H., Zakka, C., Morlon, M., Raffel, C., von Werra, L., and Wolf, T. (2025). Smollm2: When smol goes big – data-centric training of a fully open small language model. In *Proceedings of the Conference on Language Modeling*.
- Barrera, S. D. and Santos, M. J. d. (2016). Produção escrita de narrativas: influência de condições de solicitação. *Educar em Revista*, (62):69–85.
- Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.
- Ferreira, R., Freitas, E., Cabral, L., Dawn, F., Rodrigues, L., Rakovic, M., Raniel, J., and Gasevic, D. (2024). Words of wisdom: A journey through the realm of nlp for learning analytics – a systematic literature review. *Journal of Learning Analytics*, 11(3):82–105.
- Gemma Team (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Graham, S. and Harris, K. R. (2019). Evidence-based practices in writing. In Graham, S., MacArthur, C. A., and Hebert, M., editors, *Best Practices in Writing Instruction*. The Guilford Press, 3 edition.
- Hou, Z., Ciuba, A., and Li, X. (2025). Improving llm-based automatic essay scoring with linguistic features. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 41–65. PMLR.

- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Hugging Face TB (2024). Smollm-135m model card. <https://huggingface.co/HuggingFaceTB/SmolLM-135M>. Accessed: 2026-05-30.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.
- Liew, P. Y. and Tan, I. K. T. (2024). On automated essay grading using large language models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*.
- Liu, Z., Zhao, C., Iandola, F., Lai, C., Tian, Y., Fedorov, I., Xiong, Y., Chang, E., Shi, Y., Krishnamoorthi, R., Lai, L., and Chandra, V. (2024). Mobilellm: Optimizing sub-billion parameter language models for on-device use cases. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32431–32454. PMLR.
- Lobo, J., Anthony, L., Falcão, A., Xavier, C., Torrezão, N., Isotani, S., Pinto, I. I. B. S., Rodrigues, L. A. L., and Mello, R. F. (2025). Automatic scoring of elementary school essays in brazilian portuguese with llms: Comparing gemini, gpt-4o, claude, and mistral. In *Anais do Simpósio Brasileiro de Informática na Educação*.
- Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., and Lane, N. D. (2024). Small language models: Survey, measurements, and insights. *arXiv preprint arXiv:2409.15790*.
- Mello, R. F., Oliveira, H., Wenceslau, M., Batista, H., Cordeiro, T., Bittencourt, I. I., and Isotani, S. (2024). Propor’24 competition on automatic essay scoring of portuguese narrative essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 1–5.
- Nguyen, C. V., Shen, X., Aponte, R., Xia, Y., Basu, S., Hu, Z., Chen, J., Parmar, M., Kunapuli, S., Barrow, J., Wu, J., Singh, A., Wang, Y., Gu, J., Ahmed, N. K., Lipka, N., Zhang, R., Chen, X., Yu, T., Kim, S., Deilamsalehy, H., Park, N., Rimer, M., Zhang, Z., Yang, H., Mathur, P., Wu, G., Dernoncourt, F., Rossi, R. A., and Nguyen, T. H. (2025). A survey on small language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing*, pages 807–821. INCOMA Ltd.
- Oliveira, H., Mello, R. F., Miranda, P., Batista, H., Silva Filho, M. W. d., Cordeiro, T., Pinto, I. I. B. S., and Isotani, S. (2025). A benchmark dataset of narrative student essays with multi-competency grades for automatic essay scoring in brazilian portuguese. *Data in Brief*, 60:111526.
- Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.
- Qwen Team (2024). Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

- Seßler, K., Fürstenberg, M., Bühler, B., and Kasneci, E. (2025). Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the 15th International Conference on Learning Analytics and Knowledge*, pages 462–472. Association for Computing Machinery.
- Shermis, M. D. and Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge.
- Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10):1380–1400.
- Silva Filho, M. W. d. and collaborators (2024). Brazilian portuguese narrative essays dataset. <https://www.kaggle.com/datasets/moesiof/portuguese-narrative-essays>. Accessed: 2026-05-30.
- Silva Filho, M. W. d., Nascimento, A., Miranda, P., Rodrigues, L. A. L., Cordeiro, T., Isotani, S., Pinto, I. I. B. S., and Mello, R. F. (2023). Automated formal register scoring of student narrative essays written in portuguese. In *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 1–11. SBC.
- Slade, S. and Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, 57(10):1510–1529.
- Wang, D. and Wang, J. (2025). The impact mechanism of aes on improving english writing achievement. *Scientific Reports*, 15:3928.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2026). A survey of large language models. *Frontiers of Computer Science*, 20:2012627.