

Avaliação Automatizada de Feedbacks Gerados por LLMs em Questões de Biologia: Uma Abordagem via LLM-as-Judge

Ana Raissa Nascimento Silva¹, Anderson Pinheiro Cavalcanti^{1,2},
Rafael Ferreira Mello^{1,2}

¹ Cesar School

²Universidade Federal Rural de Pernambuco (UFRPE)

anaraissan@gmail.com

{anderson.pinheiro, rafael.mello}@ufrpe.br

Resumo. *Oferecer feedback formativo em larga escala é um desafio, sobretudo em disciplinas factualmente densas como a Biologia. Este trabalho aplica uma abordagem LLM-as-judge a feedbacks gerados por três Large Language Models (LLMs), GPT-5.2, Gemini-2.5-Pro e Llama3.2:3B, sob três estratégias de prompt, em 50 respostas curtas de Biologia. Dois LLMs classificaram os 450 feedbacks em cinco dimensões de conteúdo. Os modelos diferiram significativamente, mas com efeitos pequenos: o GPT-5.2 mostrou-se analítico-crítico; o Gemini-2.5-Pro, motivacional; o Llama3.2:3B, fraco em alinhamento com objetivos. Identificou-se ainda viés de autoavaliação no GPT-5.2 e concordância apenas moderada entre avaliadores, evidenciando limites da avaliação automatizada por um único LLM.*

Abstract. *Providing formative feedback at scale is a challenge, especially in factually dense disciplines such as Biology. This work applies an LLM-as-judge approach to feedbacks generated by three Large Language Models (LLMs), GPT-5.2, Gemini-2.5-Pro, and Llama3.2:3B, under three prompt strategies, on 50 short Biology responses. Two LLMs classified the 450 feedbacks across five content dimensions. The models differed significantly, but with small effect sizes: GPT-5.2 proved analytical-critical; Gemini-2.5-Pro, motivational; Llama3.2:3B, weak in goal alignment. A self-evaluation bias was identified for GPT-5.2, and inter-rater agreement was only moderate, highlighting the limits of automated evaluation by a single LLM.*

1. Introdução

O feedback é amplamente reconhecido como um dos elementos mais influentes para a aprendizagem, com efeitos diretos sobre o engajamento, a autorregulação e o desempenho dos estudantes [Hattie and Timperley 2007, Nicol and Macfarlane-Dick 2006]. No entanto, oferecer feedback formativo individualizado e de qualidade em larga escala permanece um desafio recorrente no contexto educacional brasileiro, sobretudo em razão das elevadas relações aluno-professor e do tempo demandado pela correção detalhada de produções abertas [Mello et al. 2024, Pimentel et al. 2025, Cavalcanti et al. 2021b].

O ensino de Biologia ilustra particularmente bem esse desafio. Questões dissertativas nessa área demandam não apenas a verificação de conteúdos factuais, como no-

menclatura, processos celulares ou ciclos biogeoquímicos, mas também a articulação argumentativa entre conceitos e sua contextualização em situações-problema. A correção manual de respostas abertas em Biologia é trabalhosa, suscetível a inconsistências entre avaliadores e raramente acompanhada de feedback formativo detalhado, o que motiva a busca por soluções tecnológicas capazes de viabilizar mediação pedagógica em escala [Bewersdorff et al. 2025, Cohn et al. 2024].

Os *Large Language Models* (LLMs) têm se destacado como ferramentas promissoras para a geração automática de feedback personalizado em domínios diversos [Mello et al. 2024, Yan et al. 2024]. Trabalhos recentes no contexto brasileiro investigaram seu uso para feedback em programação [Silva and da Silva Aranha 2025, Pereira and Ferreira Mello 2025] e em redações no formato ENEM [Xavier et al. 2025], evidenciando potencial significativo quanto à coerência e à adequação pedagógica das mensagens geradas. Estudos internacionais em ciências naturais, incluindo Biologia, também sugerem que LLMs podem atingir desempenho comparável ao de professores na avaliação e no fornecimento de feedback para respostas dissertativas [Bewersdorff et al. 2025, Lee et al. 2025]. Contudo, ainda existem preocupações relevantes: LLMs podem produzir conteúdo incorreto, respostas genéricas que pouco contribuem para a aprendizagem ou feedback desprovido de fundamentação pedagógica consolidada [Qian et al. 2025, Yan et al. 2024].

Para mitigar tais limitações, parte da literatura tem defendido a engenharia de prompts deliberadamente fundamentada em teorias consolidadas de feedback, como os sete princípios de boas práticas de Nicol e Macfarlane-Dick [Nicol and Macfarlane-Dick 2006]. Apesar do progresso nessa direção, a avaliação da qualidade dos feedbacks gerados sob esse paradigma tem se apoiado majoritariamente em julgamento humano, sendo uma abordagem custosa, demorada e de difícil escalabilidade [Cavalcanti et al. 2021a]. Recentemente, [Qian et al. 2025] propuseram uma abordagem *Dean of LLM Tutors*, na qual um LLM avaliador analisa, de forma automatizada e abrangente, feedbacks gerados por outros LLMs ao longo de múltiplas dimensões de qualidade. No entanto, essa abordagem foi validada exclusivamente em tarefas de Ciência da Computação em língua inglesa, restando em aberto sua aplicabilidade a outros domínios disciplinares, idiomas e níveis educacionais.

Diante dessa lacuna, este trabalho investiga a viabilidade de empregar uma abordagem de *LLM-as-judge* para avaliar automaticamente a qualidade de feedbacks gerados por LLMs em respostas curtas de Biologia. Para tanto, foram gerados feedbacks por meio de três LLMs com perfis distintos quanto ao porte e à abertura, abrangendo um modelo proprietário de grande porte, um modelo *open-source* de pequeno porte e um modelo de fronteira, todos instruídos por prompts fundamentados nas teorias de Nicol e Macfarlane-Dick [Nicol and Macfarlane-Dick 2006] e de Hattie e Timperley [Hattie and Timperley 2007]. Em seguida, dois LLMs avaliadores classificam os feedbacks ao longo das cinco dimensões de conteúdo definidas na abordagem proposta por [Qian et al. 2025]. Espera-se que os resultados contribuam tanto para o avanço de práticas de avaliação automatizada de feedback educacional em escala, quanto para uma melhor compreensão das capacidades e limitações de diferentes LLMs na produção de feedback formativo no contexto de Biologia e em língua portuguesa.

1.1. Questões e Lacunas de Pesquisa

A lacuna principal endereçada neste estudo reside na falta de investigações que apliquem abordagens de avaliação automatizada de feedback gerado por LLMs em domínios distintos da Ciência da Computação. Especificamente, busca-se compreender em que medida diferentes LLMs, instruídos por prompts teoricamente fundamentados, produzem feedbacks de qualidade comparável em uma área factualmente densa como a Biologia, e até que ponto uma abordagem automatizada de avaliação consegue distinguir tais feedbacks. Para guiar a investigação, foram formuladas as seguintes questões de pesquisa:

QUESTÃO DE PESQUISA 1 (QP1):

Em que medida os feedbacks gerados por diferentes LLMs para respostas curtas de Biologia diferem em qualidade?

QUESTÃO DE PESQUISA 2 (QP2):

Em quais dimensões de conteúdo cada LLM apresenta o melhor desempenho?

2. Trabalhos Relacionados

A literatura recente sobre feedback educacional automatizado tem evoluído de soluções centradas em verificação binária, como características de *autograders* tradicionais, para abordagens capazes de oferecer mediação pedagógica mais rica, com orientação conceitual e estímulo à reflexão [Cavalcanti et al. 2021a, Mello et al. 2024]. Essa transição tem sido fortemente impulsionada pela disseminação dos *Large Language Models* (LLMs), cuja capacidade de produzir explicações contextualizadas e personalizadas ampliou as possibilidades de feedback formativo em larga escala [Yan et al. 2024].

No contexto brasileiro, parte expressiva dos estudos sobre LLMs aplicados ao feedback educacional concentra-se no ensino de programação. [Silva and da Silva Aranha 2025] propõem um framework adaptativo baseado em LLMs para avaliação automática de código e geração de feedback formativo, alcançando 74,7% de concordância com *autograders* tradicionais e 93,3% de feedbacks classificados como coerentes em uma amostra de 300 tentativas reais. [Pimentel et al. 2025], por sua vez, exploram a geração automatizada de feedback para iniciantes em programação, ilustrando o papel da engenharia de prompts no alinhamento das respostas a objetivos pedagógicos. Em revisão sistemática recente, [Pereira and Ferreira Mello 2025] sintetizam as principais aplicações de LLMs no processo de avaliação no ensino de programação, destacando avanços e desafios relativos à fidelidade pedagógica, escalabilidade e mitigação de respostas inadequadas.

Alguns estudos têm investigado o uso de LLMs para feedback em ciências naturais, incluindo a Biologia. [Bewersdorff et al. 2025] compararam feedbacks gerados por um agente baseado em LLM com aqueles elaborados por professores e especialistas em educação científica, no contexto de protocolos experimentais de Biologia, observando desempenho próximo ao humano em múltiplas dimensões de qualidade. [Cohn et al. 2024] propuseram uma abordagem baseada em *chain-of-thought prompting* para avaliar respostas dissertativas de estudantes em avaliações formativas de Ciências, evidenciando que estratégias deliberadas de instrução podem ampliar a qualidade da análise produzida pelo modelo. [Lee et al. 2025], em uma avaliação em larga escala envolvendo seis LLMs

comerciais e *open-source* em um curso de bioinformática, reportaram desempenho de classificação e qualidade de feedback comparáveis aos de monitores humanos. Esses estudos sugerem que, embora promissor, o uso de LLMs em domínios factualmente densos ainda exige configurações cuidadosas, especialmente para mitigar respostas genéricas, imprecisões factuais e desvios pedagógicos.

No contexto de avaliação automática, trabalhos recentes têm proposto avaliações multidimensionais que combinam aspectos relativos ao conteúdo, à efetividade pedagógica e à confiabilidade do feedback [Dai et al. 2024, Qian et al. 2025]. Nessa direção, [Qian et al. 2025] introduziram a abordagem *Dean of LLM Tutors*, na qual um LLM avaliador julga feedbacks gerados por outros LLMs ao longo de 16 dimensões, incluindo conteúdo e efetividade pedagógica. Os autores demonstram que um GPT-4.1 ajustado por *fine-tuning* atinge desempenho comparável ao de avaliadores humanos especialistas (acurácia de 79,8% vs. 78,3% humana), sinalizando a viabilidade da abordagem para avaliação automatizada e em larga escala de feedback educacional.

O presente trabalho posiciona-se na intersecção dessas linhas de pesquisa. Diferentemente dos estudos brasileiros anteriores, concentrados em programação [Silva and da Silva Aranha 2025, Pimentel et al. 2025] e em produção textual em outras disciplinas [Xavier et al. 2025], este estudo foca em respostas curtas de Biologia. E diferentemente do estudo original [Qian et al. 2025], este trabalho investiga a aplicabilidade da abordagem *LLM-as-judge* em um domínio disciplinar e contexto linguístico distintos, focando especificamente na avaliação de conteúdo de feedback e contribuindo para a avaliação da generalidade da técnica e para a oferta de evidências empíricas sobre quais LLMs produzem feedback de maior qualidade nesse cenário específico.

3. Método

A metodologia adotada neste estudo foi delineada para responder às questões de pesquisa apresentadas na Seção 1, garantindo o alinhamento entre os objetivos investigativos e as estratégias de geração e avaliação automatizada de feedback. O desenho experimental organiza-se em quatro etapas: (i) seleção do conjunto de dados; (ii) geração de feedback variando três modelos LLM e três estratégias de *prompt*; (iii) avaliação automatizada dos feedbacks gerados por meio de dois LLMs atuando como avaliadores (*LLM-as-judge*); e (iv) análise comparativa dos resultados.

3.1. Conjunto de Dados

Utilizou-se o conjunto de dados PT_ASAG [Galhardi et al. 2020], um dos poucos conjuntos públicos em língua portuguesa para a tarefa de *Automatic Short Answer Grading* (ASAG). O *dataset* contempla 7.473 respostas dissertativas curtas, produzidas por 659 estudantes em 15 questões de Biologia, em português brasileiro, e disponibiliza, para cada registro, o enunciado da questão, a resposta esperada (referência) e a nota atribuída por especialistas humanos. A partir desse conjunto, selecionou-se uma amostra de 50 respostas, distribuídas entre as questões originais de modo a preservar a diversidade conceitual do *dataset*: o conjunto completo abrange 13 conceitos de Biologia, e cada um deles teve pelo menos um exemplo selecionado na amostra. Adicionalmente, a amostra preservou a distribuição original de notas atribuídas por especialistas humanos, sendo 42% de nota 0,

26% de nota 1, 18% de nota 2 e 14% de nota 3, o que assegura representatividade entre respostas integralmente incorretas e respostas integralmente corretas.

3.2. Geração de Feedback

A etapa de geração de feedback foi estruturada como um experimento fatorial 3×3 , combinando três modelos LLM com três estratégias de *prompt*, resultando em nove condições experimentais aplicadas às 50 respostas selecionadas, totalizando 450 mensagens de feedback geradas.

3.2.1. Modelos Utilizados

Foram selecionados três modelos com perfis representativos do estado da arte recente, contrastando porte, origem e abertura: (1) **GPT-5.2**¹: modelo proprietário da OpenAI, lançado em Dezembro de 2025; (2) **Gemini-2.5-Pro**²: modelo proprietário da Google, lançado em Março de 2025; e (3) **Llama3.2:3B**³: modelo *open-source* de pequeno porte (3 bilhões de parâmetros), executado localmente via *Ollama*⁴, representativo de cenários com baixo custo computacional e maior controle sobre privacidade de dados. Para todos os modelos, a temperatura foi fixada em zero, com o objetivo de reduzir a variabilidade entre execuções e favorecer a reprodutibilidade.

A seleção desses três modelos seguiu o princípio de contrastar fornecedor (OpenAI, Google e Meta), porte (de fronteira a pequeno) e grau de abertura (proprietário *versus open-source*), abordagem alinhada a estudos comparativos recentes de LLMs aplicados à avaliação e ao feedback educacional, inclusive em língua portuguesa [Lobo et al. 2025, Oketch et al. 2025]. A inclusão de um modelo *open-source* de pequeno porte é particularmente pertinente ao contexto educacional brasileiro: além de viabilizar execução local, com maior controle sobre a privacidade dos dados dos estudantes, modelos abertos como os da família Llama oferecem expressiva vantagem de custo e ampliam o acesso à tecnologia em regiões com menos recursos computacionais [Oketch et al. 2025]. Embora estudos anteriores indiquem que modelos proprietários tendem a apresentar maior alinhamento com avaliadores humanos em tarefas de correção [Lobo et al. 2025], evidências recentes sugerem que modelos abertos podem alcançar desempenho competitivo em determinadas tarefas de geração e avaliação textual, a um custo substancialmente menor [Oketch et al. 2025]. A estrutura de *prompt* (com *placeholders* para conceito, pergunta, resposta do aluno e resposta de referência) foi mantida idêntica entre os três modelos, de modo a isolar o efeito do modelo e da estratégia de *prompt* sobre a qualidade do feedback gerado.

3.2.2. Estratégias de Prompt

Foram empregadas três estratégias de *prompt*, todas exigindo saída em JSON estruturado contendo o texto do feedback. Os *prompts* estão descritos abaixo:

¹<https://developers.openai.com/api/docs/models/gpt-5.2>

²<https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/2-5-pro>

³<https://ollama.com/library/llama3.2:3b>

⁴<https://ollama.com/>

- **Base (sem fundamentação teórica):** solicita feedback claro, didático e não genérico, destacando acertos e oportunidades de melhoria em relação à resposta de referência. Esta condição opera como linha de base contra a qual as estratégias teoricamente fundamentadas são comparadas.
- **Nicol & Macfarlane-Dick** [Nicol and Macfarlane-Dick 2006]: instrui o modelo a aplicar os princípios de boas práticas de feedback: esclarecer metas e critérios, promover autoavaliação, fornecer feedback construtivo, incentivar autoestima e reduzir a lacuna entre o desempenho atual e o esperado.
- **Hattie & Timperley** [Hattie and Timperley 2007]: instrui o modelo a estruturar o feedback em torno das três perguntas orientadoras do modelo: Onde estou indo? (*Feed-up*, objetivo de aprendizagem), Como estou indo? (*Feedback*, diagnóstico do desempenho atual) e Onde ir depois? (*Feed-forward*, próximos passos para a melhoria).

3.3. Avaliação Automatizada via LLM-as-Judge

A avaliação da qualidade dos 450 feedbacks gerados foi conduzida por meio da abordagem *LLM-as-judge*, inspirada na proposta de [Qian et al. 2025]. Para investigar a robustez da avaliação automatizada e mitigar possíveis vieses associados ao uso de um único avaliador, dois LLMs foram empregados como auditores independentes: GPT-5.2 e Gemini-2.5-Pro. Cada um dos 450 feedbacks foi avaliado pelos dois modelos, totalizando 900 avaliações. É importante mencionar que essa configuração inclui tanto avaliações cruzadas (nas quais o avaliador é distinto do gerador) quanto avaliações em que o avaliador coincide com o gerador (auto-avaliações). Esse arranjo viabiliza, adicionalmente, análises sobre concordância entre avaliadores e eventuais vieses de auto-avaliação, aspectos relevantes para a fundamentação de futuras práticas de *LLM-as-judge*.

Os LLMs avaliadores receberam, para cada feedback, instruções para classificá-lo ao longo das cinco dimensões de conteúdo da abordagem de [Qian et al. 2025], adaptadas para o contexto deste estudo:

- **Alinhamento com objetivos:** em que medida o feedback conecta os comentários ao objetivo de aprendizagem da questão;
- **Especificidade:** em que medida o feedback explica o que e como melhorar, oferecendo exemplos concretos;
- **Tom motivacional:** em que medida o feedback adota linguagem encorajadora e respeitosa;
- **Reconhecimento de pontos fortes:** em que medida o feedback identifica claramente o que foi feito adequadamente pelo estudante;
- **Orientação sobre pontos fracos:** em que medida o feedback identifica o que precisa melhorar e oferece orientação prática para a melhoria.

Embora o estudo de [Qian et al. 2025] proponha um conjunto maior de dimensões, optou-se, devido a limitações de recursos para utilização de modelos proprietários, por utilizar somente as dimensões de conteúdo neste trabalho. Cada dimensão foi pontuada em uma escala Likert de três pontos (1 = fraco, 2 = médio, 3 = forte). A pontuação geral de cada feedback corresponde à média simples das cinco dimensões, posteriormente classificada em três categorias: Fraco (1,0–1,5), Médio (1,6–2,3) e Forte (2,4–3,0). O LLM avaliador foi instruído a retornar, para cada dimensão, a pontuação atribuída e uma justificativa textual sucinta, em formato JSON estruturado, viabilizando análises posteriores da

fundamentação de cada decisão. A temperatura do modelo avaliador também foi fixada em zero.

3.4. Análise dos Resultados

Para responder à **QP1**, compararam-se as distribuições dos escores Likert agregados (média das cinco dimensões) entre os três modelos geradores por meio do teste de Kruskal-Wallis, seguido de comparações *post-hoc* pareadas (Mann-Whitney com correção de Bonferroni). Para a **QP2**, conduziu-se análise descritiva por dimensão, possibilitando identificar em quais aspectos do feedback cada modelo e cada estratégia de *prompt* apresenta desempenho relativamente superior ou inferior. Complementarmente, quantificou-se a concordância entre os dois avaliadores por meio da correlação intraclassa (ICC), do Kappa de Cohen ponderado (para a classificação em três níveis) e da correlação de Spearman; e avaliou-se o viés de auto-avaliação por meio do teste de Wilcoxon pareado. As análises foram realizadas em Python, utilizando as bibliotecas `pandas`, `scipy`, `pingouin` e `scikit-learn`.

4. Resultados

Esta seção apresenta os resultados da avaliação automatizada dos 450 feedbacks gerados, totalizando 900 avaliações (450 feedbacks $\times$ 2 avaliadores). Os resultados estão organizados em quatro partes: (i) visão geral dos escores por condição (Subseção 4.1); (ii) comparação entre modelos geradores (QP1) (Subseção 4.2); (iii) efeito das estratégias de *prompt* e análise da concordância entre avaliadores (Subseção 4.3); e (iv) análise por dimensão de conteúdo (QP2) (Subseção 4.4).

4.1. Visão Geral dos Escores

A Tabela 1 apresenta o escore médio das cinco dimensões de conteúdo, em escala Likert de 1 a 3, agregado sobre as 50 respostas avaliadas, para cada uma das 18 condições experimentais (3 modelos geradores $\times$ 3 estratégias de *prompt* $\times$ 2 avaliadores). Cada célula original (avaliador específico) é baseada em $n = 50$; as colunas *Média aval.* agregam $n = 100$ (*pooled* entre os dois avaliadores); a linha *Média* agrega $n = 150$ por célula de avaliador e $n = 300$ em sua interseção com *Média aval.*. Células sombreadas indicam pares de autoavaliação, em que o LLM gerador coincide com o LLM avaliador. Em negrito, a melhor estratégia de *prompt* por gerador e por estratégia de *prompt*.

Tabela 1. Escore médio das cinco dimensões de conteúdo em escala Likert (1–3), no formato *M (DP)*, por modelo gerador, estratégia de *prompt* e avaliador.

Gerador	Avaliador: GPT-5.2			Avaliador: Gemini-2.5-Pro			Média aval.		
	Base	Hattie	Nicol	Base	Hattie	Nicol	Base	Hattie	Nicol
GPT-5.2	2,32 (0,26)	2,34 (0,15)	2,52 (0,24)	2,20 (0,36)	2,24 (0,37)	2,20 (0,37)	2,26 (0,32)	2,29 (0,29)	2,36 (0,35)
Gemini-2.5-Pro	2,06 (0,44)	2,26 (0,23)	2,19 (0,37)	2,02 (0,53)	2,28 (0,39)	2,30 (0,49)	2,04 (0,49)	2,27 (0,32)	2,24 (0,44)
Llama3.2:3B	2,25 (0,39)	1,94 (0,24)	1,92 (0,30)	2,19 (0,35)	2,17 (0,37)	2,21 (0,34)	2,22 (0,37)	2,05 (0,33)	2,07 (0,35)
Média	2,21 (0,38)	2,18 (0,27)	2,21 (0,39)	2,14 (0,43)	2,23 (0,38)	2,24 (0,40)	2,17 (0,41)	2,21 (0,33)	2,22 (0,40)

Os escores variaram entre 1,92 e 2,52, situando a maioria dos feedbacks na categoria *Medium* (1,6–2,3), com a configuração *GPT-5.2 + prompt Nicol*, sob auto-avaliação, atingindo a categoria *Strong* ($\geq 2,4$). O menor escore médio observado foi o do par *Llama3.2:3B + prompt Nicol*, sob avaliação do GPT-5.2 (1,92). Considerando a média entre os dois avaliadores (três últimas colunas da Tabela 1), a estratégia de *prompt* mais

eficaz variou conforme o gerador: Nicol para o GPT-5.2 (2,36), Hattie para o Gemini-2.5-Pro (2,27) e Base para o Llama3.2:3B (2,22). Já na média entre os três modelos geradores (última linha), as estratégias teoricamente fundamentadas apresentaram leve vantagem sobre a linha de base (Nicol 2,22 e Hattie 2,21, contra 2,17 do *prompt* Base).

4.2. Comparação entre Modelos Geradores

Para responder à **QP1**, compararam-se os escores agregados (média das cinco dimensões) entre os três modelos geradores, considerando as 300 avaliações de cada modelo (3 *prompts* × 50 respostas × 2 avaliadores). O GPT-5.2 obteve o maior escore médio ($M = 2,31$, $DP = 0,32$), seguido pelo Gemini-2.5-Pro ($M = 2,18$, $DP = 0,43$) e pelo Llama3.2:3B ($M = 2,11$, $DP = 0,36$). O teste de Kruskal-Wallis indicou diferença global estatisticamente significativa entre os modelos ($H = 48,69$, $p < 0,001$).

As comparações *post-hoc* (Mann-Whitney com correção de Bonferroni) confirmaram que todas as três diferenças par a par são estatisticamente significativas: GPT-5.2 vs. Gemini-2.5-Pro ($p = 0,010$; rank-biserial $r = 0,14$), GPT-5.2 vs. Llama3.2:3B ($p < 0,001$; $r = 0,33$) e Gemini-2.5-Pro vs. Llama3.2:3B ($p < 0,001$; $r = 0,18$). Contudo, é importante destacar que os tamanhos de efeito variam de pequenos a moderados ($0,14 \leq r \leq 0,33$): embora a hierarquia GPT-5.2 > Gemini-2.5-Pro > Llama3.2:3B seja estatisticamente robusta, as diferenças absolutas na escala Likert são modestas. Em particular, o Llama3.2:3B, modelo de apenas 3 bilhões de parâmetros, executável localmente, situa-se a 0,07 ponto do Gemini-2.5-Pro, um modelo proprietário de fronteira. Ainda que essa diferença seja significativa, sua pequena magnitude sugere que modelos abertos e de baixo custo podem produzir feedback de qualidade próxima à de modelos de fronteira em tarefas estruturadas, o que é relevante para cenários educacionais com restrições de custo computacional e de privacidade de dados.

4.3. Efeito das Estratégias de Prompt e Concordância entre Avaliadores

A análise por estratégia de *prompt* revela padrões distintos para cada modelo gerador. Para o GPT-5.2, os escores são relativamente uniformes entre as três estratégias quando avaliado pelo Gemini-2.5-Pro (2,20–2,24), mas o *prompt* fundamentado em Nicol e Macfarlane-Dick se destaca quando avaliado pelo próprio GPT-5.2 (2,52). Para o Gemini-2.5-Pro, os *prompts* fundamentados em Hattie e Timperley (2,26–2,28) e em Nicol e Macfarlane-Dick (2,19–2,30) consistentemente superam o *prompt* Base (2,02–2,06) em ambos os avaliadores. Para o Llama3.2:3B, observa-se padrão oposto na avaliação pelo GPT-5.2: o *prompt* Base obteve o maior escore (2,25), enquanto Hattie (1,94) e Nicol (1,92) ficaram substancialmente abaixo. Esse comportamento sugere que *prompts* teoricamente fundamentados, ao adicionarem instruções mais complexas e longas, podem ultrapassar a capacidade de seguimento de instruções de modelos menores, levando à degradação da saída em vez do seu aprimoramento.

A concordância entre os dois avaliadores, contudo, mostrou-se apenas moderada a baixa. A correlação entre os escores atribuídos por GPT-5.2 e Gemini-2.5-Pro a um mesmo feedback foi fraca a moderada (Spearman $\rho = 0,34$; Pearson $r = 0,42$; $p < 0,001$), e o coeficiente de correlação intraclass indicou confiabilidade apenas razoável para um único avaliador ($ICC(1,1) = 0,42$; $IC_{95\%} [0,34; 0,49]$). A concordância na classificação em três níveis (Fraco/Médio/Forte) foi baixa (Kappa de Cohen ponderado = 0,20). Notavelmente, a diferença média absoluta entre os dois avaliadores

para um mesmo feedback foi de 0,32 ponto, magnitude superior às próprias diferenças observadas entre os modelos geradores (0,07 a 0,19 ponto). Esse achado constitui uma limitação importante da abordagem *LLM-as-judge* tal como aplicada: a variância introduzida pela escolha do avaliador supera a variância entre os objetos avaliados, e a concordância obtida é substancialmente inferior à reportada entre avaliadores humanos especialistas no estudo original de [Qian et al. 2025] (Fleiss' Kappa = 0,77). Isso sugere que, dependendo do contexto avaliado e da língua, a avaliação automatizada por um único LLM ainda carece de confiabilidade suficiente, reforçando a necessidade de múltiplos avaliadores em paralelo e de validação humana.

O teste de Wilcoxon pareado, aplicado às notas que cada feedback recebeu em autoavaliação e em avaliação cruzada, identificou viés estatisticamente significativo apenas para o GPT-5.2: o modelo pontuou os próprios feedbacks em média 0,18 ponto acima do avaliador externo ($W = 1620$, $p < 0,001$). Para o Gemini-2.5-Pro, o viés foi pequeno e não significativo (+0,03 ponto; $W = 2986$, $p = 0,16$). Esses resultados corroboram preocupações da literatura sobre vieses em LLMs avaliadores [Qian et al. 2025] e reforçam a recomendação de não empregar um LLM como avaliador de seus próprios feedbacks, privilegiando avaliações cruzadas e o uso de múltiplos avaliadores.

4.4. Análise por Dimensão de Conteúdo

Para responder à **QP2**, analisaram-se os escores Likert decompostos pelas cinco dimensões de conteúdo, agregados por modelo gerador e por estratégia de *prompt*. A Figura 1 apresenta essa decomposição em dois painéis: no painel (a), o escore médio de cada dimensão por modelo gerador (agregado entre *prompts* e avaliadores); no painel (b), o escore médio por estratégia de *prompt* (agregado entre modelos e avaliadores).

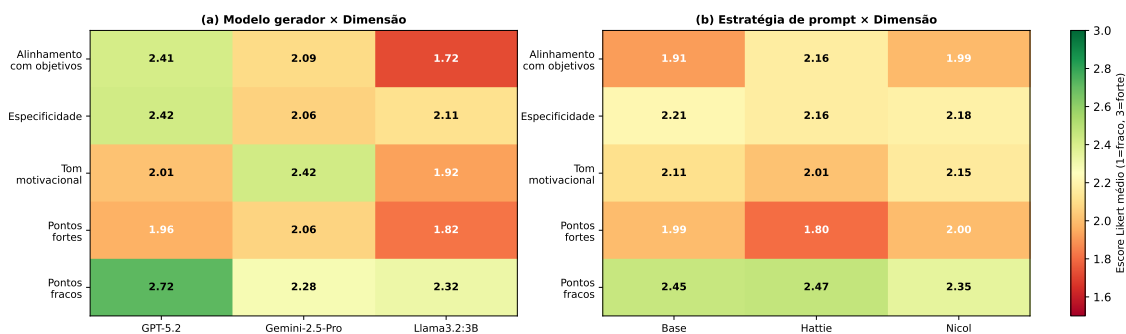


Figura 1. Escoré Likert médio (1–3) por dimensão de conteúdo: (a) por modelo gerador; (b) por estratégia de *prompt*.

A análise por modelo gerador revela que cada LLM exibe um perfil distinto de pontos fortes e fracos ao longo das dimensões. O GPT-5.2 obteve os maiores escores nas dimensões de *Pontos fracos* (2,72), *Especificidade* (2,42) e *Alinhamento com objetivos* (2,41), mas teve desempenho notavelmente inferior em *Tom motivacional* (2,01) e *Pontos fortes* (1,96). Esse padrão sugere um perfil mais analítico e crítico, eficaz em identificar lacunas e fornecer orientação específica, mas menos efetivo no reconhecimento explícito de acertos e no uso de linguagem encorajadora. O Gemini-2.5-Pro apresentou o perfil oposto: seu maior escore foi na dimensão de *Tom motivacional* (2,42), substancialmente acima dos demais modelos, com desempenho relativamente equilibrado nas

demais dimensões (2,06–2,28). Isso indica que o Gemini-2.5-Pro tende a gerar feedback de natureza afetivamente mais acolhedora. Já o Llama3.2:3B obteve o pior desempenho na dimensão de *Alinhamento com objetivos* (1,72), o menor escore observado em toda a matriz; e também desempenho inferior em *Tom motivacional* (1,92) e *Pontos fortes* (1,82). Por outro lado, manteve desempenho razoável em *Pontos fracos* (2,32) e *Especificidade* (2,11). Esse padrão sugere que o modelo, de menor porte, é capaz de identificar e descrever problemas concretos nas respostas, mas tem dificuldade em conectá-los explicitamente aos objetivos de aprendizagem da questão.

A análise por estratégia de *prompt* (painel b) revela diferenças mais discretas, mas com padrões interpretáveis. O *prompt* fundamentado em Hattie e Timperley produziu o maior escore em *Alinhamento com objetivos* (2,16, contra 1,91 do *prompt* Base), o que é consistente com a estrutura do modelo de Hattie, cuja componente *Feed-up* explicitamente orienta o feedback aos objetivos de aprendizagem. Por outro lado, esse mesmo *prompt* obteve o menor escore em *Pontos fortes* (1,80), provavelmente porque a estrutura dessa teoria concentra-se em diagnóstico e próximos passos, deixando menor espaço para reconhecimento explícito de acertos. O *prompt* fundamentado em Nicol e Macfarlane-Dick resultou no perfil mais equilibrado entre dimensões (1,99–2,35), o que é coerente com sua maior amplitude conceitual, abrangendo simultaneamente metas, autoavaliação, motivação e redução de lacuna de desempenho. O *prompt* Base, embora menos consistente em *Alinhamento*, obteve escore razoavelmente competitivo em *Pontos fracos* (2,45) e *Especificidade* (2,21).

Em síntese, a análise dimensional indica que: (i) nenhum modelo é uniformemente superior em todas as dimensões, sendo a escolha do LLM gerador dependente do tipo de feedback pedagogicamente prioritário; (ii) a engenharia de *prompt* fundamentada em teoria pode, em dimensões específicas (notadamente *Alinhamento*), trazer ganhos mensuráveis em relação a *prompts* não fundamentados; e (iii) a restrição estrutural imposta ao formato do feedback (duas sentenças curtas) cria um piso comum de fragilidade na dimensão de *Pontos fortes*, sugerindo uma direção de aprimoramento via flexibilização do tamanho da saída.

5. Discussão

Este estudo investigou a viabilidade de uma abordagem *LLM-as-judge* para avaliar automaticamente feedbacks gerados em respostas curtas de Biologia em português, comparando três LLMs com perfis distintos quanto a porte e abertura. Os achados indicam que a abordagem é viável e revela padrões interpretáveis, mas que sua aplicação no domínio em estudo requer cautela considerável. Esta seção interpreta os resultados em diálogo com a literatura recente e discute suas implicações pedagógicas e metodológicas.

A hierarquia observada entre os modelos geradores (GPT-5.2 > Gemini-2.5-Pro > Llama3.2:3B), estatisticamente significativa em todas as comparações par a par, alinha-se à expectativa de que modelos proprietários de fronteira tendem a produzir feedback de qualidade ligeiramente superior, em consonância com achados de [Lobo et al. 2025] para redações no contexto brasileiro. Contudo, os tamanhos de efeito pequenos a moderados sugerem que a hierarquia, embora real, é de magnitude prática limitada. Em particular, a diferença de apenas 0,07 ponto entre o Llama3.2:3B (3 bilhões de parâmetros, executável localmente) e o Gemini-2.5-Pro (modelo proprietário de fronteira) reforça a argumentação

de [Oketch et al. 2025] sobre a competitividade emergente de modelos abertos, especialmente quando custo computacional, controle local e privacidade dos dados dos estudantes constituem requisitos relevantes, como tipicamente ocorre em instituições educacionais brasileiras com recursos limitados.

Mais importante que a ordenação dos modelos, a análise dimensional revelou que cada LLM exibe um perfil pedagógico distinto. O GPT-5.2 mostrou-se analítico-crítico, eficaz em identificar lacunas e fornecer orientação específica, mas notavelmente fraco no reconhecimento explícito de pontos fortes e no uso de tom encorajador. O Gemini-2.5-Pro apresentou perfil oposto, destacando-se no tom motivacional. O Llama3.2:3B demonstrou maior dificuldade em conectar suas observações aos objetivos de aprendizagem da questão. Esses perfis sugerem que a escolha do modelo deve ser informada pelo objetivo pedagógico do feedback: para estudantes em fase de diagnóstico e correção, o GPT-5.2 pode ser mais adequado; para estudantes que necessitam de suporte motivacional, o Gemini-2.5-Pro pode ser preferível. Essa perspectiva ressoa com a constatação de [Xavier et al. 2025] de que não existe modelo universalmente superior, e amplia o debate sobre personalização do feedback para incluir a personalização do próprio gerador conforme a finalidade formativa.

A análise das estratégias de *prompt* revelou um padrão particularmente relevante para a engenharia de *prompts* em educação: enquanto os modelos de maior porte se beneficiaram, em diferentes graus, das estratégias teoricamente fundamentadas (Hattie e Timperley melhorou o Alinhamento com objetivos para o Gemini-2.5-Pro; Nicol e Macfarlane-Dick produziu o maior escore agregado para o GPT-5.2), o Llama3.2:3B sofreu queda substancial de desempenho com esses mesmos *prompts* (de 2,25 com *prompt* Base para 1,94 e 1,92 com Hattie e Nicol, respectivamente, na avaliação cruzada pelo GPT-5.2). Esse achado sugere que *prompts* mais longos e estruturados podem ultrapassar a capacidade de seguimento de instruções de modelos pequenos, levando à degradação da saída em vez de seu aprimoramento. Para profissionais da educação que considerem soluções locais com modelos abertos de pequeno porte, essa é uma evidência relevante: a engenharia de *prompts* elaborada não deve ser adotada acriticamente, e *prompts* mais simples podem ser preferíveis para esses modelos.

Do ponto de vista metodológico, dois achados merecem atenção. Primeiro, a concordância apenas moderada entre os dois LLMs avaliadores ($ICC = 0,42$; Kappa ponderado = 0,20), com diferença média de 0,32 ponto entre eles, magnitude superior às próprias diferenças observadas entre os modelos geradores (0,07 a 0,19 ponto), constitui uma limitação importante da abordagem. Em comparação, o estudo original de [Qian et al. 2025] reportou concordância humana de Fleiss' Kappa = 0,77. Mesmo que nosso estudo não tenha utilizado todas as dimensões propostas por [Qian et al. 2025], essa disparidade sinaliza que, dependendo do contexto, a confiabilidade da avaliação automatizada por um único LLM ainda é insuficiente para substituir o julgamento humano, e indica a importância de protocolos com múltiplos avaliadores em paralelo. Segundo, o viés sistemático de auto-avaliação observado para o GPT-5.2 (+0,18 ponto, $p < 0,001$), e ausente no Gemini-2.5-Pro (+0,03 ponto, n.s.), refina o entendimento sobre vieses em *LLM-as-judge*: o viés não é uniforme entre modelos. Notavelmente, o GPT-5.2 também apresentou os menores desvios padrão entre suas autoavaliações (entre 0,15 e 0,26 contra 0,36–0,37 quando avaliado pelo Gemini-2.5-Pro), sugerindo que sua inflação na autoavaliação se

associa a uma maior consistência interna. Esse padrão reforça a recomendação prática de evitar configurações em que o gerador e o avaliador coincidem.

6. Limitações

O trabalho em questão possui algumas limitações. A principal delas é a ausência de validação humana das avaliações automatizadas: embora a concordância entre os dois LLMs avaliadores tenha sido quantificada, ela representa convergência entre máquinas, e não validade pedagógica frente a especialistas humanos. Adicionalmente, a amostra é restrita a 50 questões de um único conjunto (PT_ASAG), de um único nível educacional (8º ano do ensino fundamental). Há também limitações de escopo metodológico. Outro ponto importante foi a decisão de utilizar somente as dimensões de conteúdo da abordagem de [Qian et al. 2025], deixando de lado as dimensões de efetividade pedagógica e de detecção de alucinação, onde esta última é particularmente relevante em um domínio factualmente denso como a Biologia. Por fim, as 50 questões foram reutilizadas em nove condições experimentais, o que implica certa dependência entre observações que não foi formalmente modelada nas análises estatísticas.

7. Considerações Finais

Este trabalho investigou a viabilidade de uma abordagem *LLM-as-judge* para avaliar automaticamente feedbacks gerados por LLMs em respostas curtas de Biologia em língua portuguesa, contribuindo para a transferência da técnica proposta por [Qian et al. 2025] para um domínio disciplinar, linguístico e educacional ainda não explorado. Em resposta à **QP1**, os resultados revelaram uma hierarquia estatisticamente significativa entre os modelos geradores (GPT-5.2 > Gemini-2.5-Pro > Llama3.2:3B), porém com tamanhos de efeito pequenos a moderados, evidenciando que modelos de fronteira ainda apresentam vantagem na qualidade do feedback, mas que essa vantagem é de magnitude limitada e pode não justificar, em todos os cenários, o custo computacional adicional. Em resposta à **QP2**, a análise dimensional demonstrou que cada modelo exibe um perfil pedagógico distinto, analítico-crítico para o GPT-5.2, motivacional para o Gemini-2.5-Pro e factualmente orientado, ainda que limitado em alinhamento com objetivos, para o Llama3.2:3B.

As principais contribuições do estudo são: (i) a primeira aplicação, no nosso conhecimento, da abordagem *LLM-as-judge* para avaliação de feedback educacional em Biologia em língua portuguesa, mostrando que a técnica é tecnicamente viável mas que sua confiabilidade requer cuidados adicionais nesse contexto; (ii) a caracterização empírica de perfis pedagógicos distintos entre LLMs proprietários e abertos, fornecendo critérios práticos para a seleção do modelo em função do objetivo formativo; e (iii) a evidência de viés significativo de auto-avaliação no GPT-5.2 mas não no Gemini-2.5-Pro, refinando recomendações metodológicas sobre o uso de *LLM-as-judge*.

Como trabalhos futuros, três direções se destacam. A primeira é a validação humana de uma subamostra das avaliações, para estabelecer a confiabilidade dos juízes LLM contra anotadores especialistas, em protocolo inspirado em [Qian et al. 2025]. A segunda é a extensão do desenho experimental para incluir as demais dimensões da abordagem *Dean of LLM Tutors*, efetividade pedagógica e detecção de alucinação, com particular relevância para esta última no contexto de Biologia, dada a densidade factual da disciplina. A terceira é a investigação sistemática do efeito do tamanho do feedback, verificando sua relação com as dimensões de avaliação e teorias de feedback.

Cuidados Éticos

Este estudo baseia-se exclusivamente no conjunto de dados público PT_ASAG [Galhardi et al. 2020], previamente disponibilizado pela comunidade de pesquisa, contendo respostas anonimizadas de estudantes para fins de investigação em avaliação automática. Nenhuma coleta primária de dados envolvendo estudantes humanos foi conduzida no escopo deste trabalho, e nenhum identificador pessoal foi processado ou apresentado. Os feedbacks gerados pelos LLMs não foram apresentados a estudantes reais, restringindo-se ao escopo experimental. Portanto, o estudo não foi elegível para avaliação por um Comitê de Ética em Pesquisa.

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 por seu conteúdo.

Disponibilidade de artefatos

Os artefatos deste estudo estão disponíveis no link: <https://anonymous.4open.science/r/feedback-sbie2026-5F77/>.

Referências

- [Bewersdorff et al. 2025] Bewersdorff, A. et al. (2025). Towards adaptive feedback with AI: Comparing the feedback quality of LLMs and teachers on experimentation protocols. In *arXiv preprint arXiv:2502.12842*.
- [Cavalcanti et al. 2021a] Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., and Mello, R. F. (2021a). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027.
- [Cavalcanti et al. 2021b] Cavalcanti, A. P., Mello, R. F., Miranda, P., Nascimento, A., and Freitas, F. (2021b). Utilização de recursos linguísticos para classificação automática de mensagens de feedback. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 861–872. SBC.
- [Cohn et al. 2024] Cohn, C. et al. (2024). A chain-of-thought prompting approach with LLMs for evaluating students’ formative assessment responses in science. In *Proceedings of the AAAI Conference on Artificial Intelligence*. arXiv:2403.14565.
- [Dai et al. 2024] Dai, W., Tsai, Y.-S., Lin, J., Aldino, A., Jin, H., Li, T., Gašević, D., and Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, 7:100299.
- [Galhardi et al. 2020] Galhardi, L. B., Senefonte, H., and Brancher, J. D. (2020). Portuguese automatic short answer grading. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1373–1382.
- [Hattie and Timperley 2007] Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1):81–112.

- [Lee et al. 2025] Lee, P. et al. (2025). Automated assignment grading with large language models: insights from a bioinformatics course. *BMC Bioinformatics*. Available at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12261420/>.
- [Lobo et al. 2025] Lobo, J., Anthony, L., Falcão, A., Xavier, C., Torrezão, N., Isotani, S., Bittencourt, I. I., Rodrigues, 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 XXXVI Simpósio Brasileiro de Informática na Educação (SBIE 2025)*, pages 168–181.
- [Mello et al. 2024] Mello, R. F., Freitas, E., Cabral, L., Pereira, F. D., Rodrigues, L., Rakovic, M., Raniel, J., and Gasević, D. (2024). Words of wisdom: A journey through the realm of natural language processing for learning analytics—a systematic literature review. *Journal of Learning Analytics*, 11(3):82–105.
- [Nicol and Macfarlane-Dick 2006] Nicol, D. J. and Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2):199–218.
- [Oketch et al. 2025] Oketch, K., Lalor, J. P., Yang, Y., and Abbasi, A. (2025). Bridging the LLM accessibility divide? performance, fairness, and cost of closed versus open LLMs for automated essay scoring. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*. arXiv:2503.11827.
- [Pereira and Ferreira Mello 2025] Pereira, A. F. and Ferreira Mello, R. (2025). A systematic literature review on large language models applications in computer programming teaching evaluation process. *IEEE Access*, 13:113449–113460.
- [Pimentel et al. 2025] Pimentel, E. P., da Silva, L. D. L., Takeuti, R. H., and Braga, J. C. (2025). Abordagem com LLM para geração automatizada de feedback no ensino de programação de computadores. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 590–602.
- [Qian et al. 2025] Qian, K., Cheng, Y., Guan, R., Dai, W., Jin, F., Yang, K., Nawaz, S., Swiecki, Z., Chen, G., Yan, L., and Gasević, D. (2025). Dean of LLM tutors: Exploring comprehensive and automated evaluation of LLM-generated educational feedback via LLM feedback evaluators. *arXiv preprint arXiv:2508.05952*.
- [Silva and da Silva Aranha 2025] Silva, F. G. and da Silva Aranha, E. H. (2025). Feedback formativo automatizado com llms: Desenvolvimento e análise de um sistema para aprendizagem progressiva em programação. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1158–1172. SBC.
- [Xavier et al. 2025] Xavier, C., da Costa, N. T., Valdo, A. K., Alves, G., Rodrigues, L., Rodrigues, L. F., Silva, M., Neto, R., Falcão, T. P., Gasevic, D., et al. (2025). Human teacher vs. LLM-generated feedback in secondary education: A comparative study on student perceptions. In *European Conference on Technology Enhanced Learning*, pages 534–548.
- [Yan et al. 2024] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., and Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.