

Feedback humano ou artificial? Comparando percepções de estudantes sobre feedback gerado por LLM e professores em atividades de POO

Marcel Melo¹, Luiz Rodrigues², Rafael Ferreira Mello³,
Cleon Xavier⁴, Rafael Dias Araújo⁵

¹Núcleo de Computação – Instituto Federal Goiano – campus Morrinhos

²Universidade Tecnológica Federal do Paraná (UTFPR) – Campus Apucarana

³Universidade Federal Rural de Pernambuco (UFRPE)

⁴Instituto Federal Goiano – campus Iporá

⁵Faculdade de Computação – Universidade Federal de Uberlândia (UFU)

marcel.melo@ifgoiano.edu.br, luiz.r70@gmail.com, rafael.mello@ufrpe.br

cleon.junior@ifgoiano.edu.br, rafael.araujo@ufu.br

Abstract. *This study compared students' perceptions of formative feedback from the teacher and feedback generated by an LLM during an Object-Oriented Programming activity. Although the LLM-generated feedback showed slightly higher mean scores across some constructs, the quantitative analysis using the Wilcoxon test did not indicate statistically significant differences between the two types of feedback. The qualitative analysis showed that students perceived the LLM feedback as more detailed, explanatory, and structured. The results suggest that LLMs can support teaching practice, especially as a complementary tool to expand the provision of formative feedback in programming activities.*

Resumo. *Este trabalho comparou a percepção de estudantes sobre o feedback formativo elaborado pelo professor e o feedback gerado por uma LLM em uma atividade de Programação Orientada a Objetos. Apesar do feedback gerado pela LLM apresentar médias ligeiramente superiores em alguns constructos, a análise quantitativa baseada teste de Wilcoxon, não indicou diferenças estatisticamente significativas entre os dois tipos de feedback. A análise qualitativa mostrou que os estudantes perceberam o feedback da LLM como mais detalhado, explicativo e estruturado. Os resultados sugerem que as LLMs podem apoiar a prática docente, especialmente como ferramenta complementar para ampliar a oferta de feedback formativo em atividades de programação.*

1. Introdução

O feedback é reconhecido na literatura educacional como um elemento central para apoiar a aprendizagem, pois fornece informações que ajudam o estudante a compreender a distância entre seu desempenho atual e os objetivos esperados [Hattie and Timperley 2007, Shute 2008]. Nesse sentido, [Hattie and Timperley 2007]

definem feedback como informação fornecida por diferentes agentes sobre o desempenho ou a compreensão do aprendiz, enquanto [Shute 2008] o caracteriza como uma informação destinada a modificar o pensamento ou comportamento do estudante para melhorar sua aprendizagem. Assim, o feedback formativo contribui para os processos de ensino e aprendizagem ao indicar não apenas se uma resposta está correta ou incorreta, mas também explicar erros, apontar aspectos a serem aprimorados e sugerir estratégias para que o estudante avance em direção aos objetivos educacionais [Keuning et al. 2018, Mandouit and Hattie 2023, Hattie and Timperley 2007, Shute 2008].

No ensino de programação, o feedback formativo apresenta desafios específicos, pois as atividades exigem não apenas correção sintática ou execução adequada do código, mas também raciocínio lógico, depuração, organização da solução e compreensão conceitual [Lohr et al. 2024]. No contexto da Programação Orientada a Objetos (POO), por exemplo, devido à sua natureza abstrata e complexa, a avaliação deve considerar não apenas se o código foi executado corretamente, mas também se o estudante compreendeu e aplicou adequadamente os princípios da orientação a objetos envolvidos na atividade [Efan et al. 2023]. Contudo, esse processo demanda tempo e esforço docente, especialmente em turmas numerosas, o que reforça a necessidade de estratégias de apoio à avaliação e à geração de feedback formativo [Silva and Aranha 2025, Yousef et al. 2025].

Nesse cenário, ferramentas de feedback automático e os grandes modelos de linguagem (*Large Language Models*, LLMs) têm sido investigados como alternativas para apoiar o trabalho docente. Estudos sobre feedback automático em programação indicam que muitas ferramentas tradicionais concentram-se na identificação de erros, na execução de testes ou na indicação de problemas sintáticos, oferecendo menor apoio quanto a como o estudante pode corrigir suas dificuldades ou avançar conceitualmente [Keuning et al. 2018]. Os LLMs ampliam esse cenário ao possibilitar a geração de feedback em linguagem natural, potencialmente mais personalizado, explicativo e alinhado a modelos pedagógicos. Entretanto, o uso dessas tecnologias ainda demanda controle pedagógico sobre o tipo, o conteúdo e a adequação das respostas geradas [Lohr et al. 2024].

Além das investigações sobre a capacidade técnica dos LLMs de gerar feedback, estudos recentes têm analisado como os estudantes percebem esses comentários em comparação com o feedback produzido por professores. No contexto da educação básica, pesquisas observaram que estudantes não perceberam diferenças significativas de qualidade entre o feedback docente e o feedback assistido por LLM, pois muitos não conseguiram distinguir a origem do feedback [Xavier et al. 2026]. Em atividades de escrita, os resultados também indicam que o feedback gerado por Inteligência Artificial (IA) pode apresentar boa qualidade, embora o feedback docente tenha sido mais frequentemente incorporado pelos estudantes durante o processo de revisão [Yu and Xie 2025].

Em uma disciplina de programação em Java, pesquisadores identificaram que o feedback do instrutor foi percebido como mais útil do que o feedback gerado por IA, embora dimensões como justiça e encorajamento não tenham apresentado diferenças significativas [Er et al. 2025]. Vale notar que, em geral, esses estudos frequentemente geram feedback por meio de comandos que não estabelecem uma fundamentação pedagógica para orientar o processo formativo dos estudantes, ao contrário do que se estabelece na literatura da área [Hattie and Timperley 2007, Lohr et al. 2024, Shute 2008].

Apesar desses avanços, nota-se que ainda há uma lacuna na comparação das percepções dos estudantes sobre o feedback elaborado pelo professor e o feedback gerado por LLM quando este é estruturado com base em fundamentação teórica relevante, especialmente no contexto de atividades avaliativas de POO. Diante disso, este trabalho busca investigar como os estudantes percebem esses dois tipos de feedback, considerando aspectos como qualidade, clareza, utilidade e potencial formativo.

2. Trabalhos relacionados

Na literatura, diferentes modelos teóricos orientam a elaboração de feedback formativo. [Shute 2008] destaca que o feedback deve ser claro, específico, objetivo, focado na tarefa e orientado à melhoria da aprendizagem, evitando comentários vagos, julgamentos pessoais, comparações entre estudantes e excesso de informação. De forma complementar, [Nicol and Macfarlane-Dick 2006] defendem que o feedback deve promover a autorregulação, ajudando o estudante a compreender critérios de qualidade, avaliar seu próprio desempenho e identificar ações de melhoria, além de apoiar o professor no ajuste do ensino.

Entre os modelos mais influentes, [Hattie and Timperley 2007] propõem que um feedback eficaz deve responder três perguntas: “Onde estou?” (*Feed Up*), “Como estou indo?” (*Feed Back*) e “Para onde vou agora?” (*Feed Forward*). O *Feed Up* esclarece quais são os objetivos de aprendizagem e os critérios de sucesso esperados. O *Feed Back* foca em fornecer informações específicas sobre o progresso do estudante, indicando acertos, erros ou estratégias equivocadas. Por fim, o *Feed Forward* orienta o estudante sobre os próximos passos, novas estratégias, aprofundamentos ou desafios que podem promover avanços contínuos. Integrados, esses três elementos constituem um ciclo de feedback que não apenas informa, mas também amplia a autonomia e o engajamento do estudante.

Adicionalmente, o modelo de Hattie e Timperley (2007) indica que o feedback atua em quatro níveis distintos de foco: o nível da tarefa, que indica se a submissão está correta ou não; o nível do processo, que apresenta estratégias necessárias para compreender ou realizar a tarefa; o nível de autorregulação, que apresenta o monitoramento, a direção e a regulação das próprias ações do estudante; e o nível do eu/pessoal, que apresenta avaliações focadas nas características do aluno, como elogios.

Este modelo tem sido amplamente utilizado para estruturar e avaliar feedback educacional. Estudos aplicam esse modelo em diferentes contextos, como a percepção discente sobre feedback [Brooks et al. 2019], a avaliação de feedback escrito em trabalhos de conclusão no ensino superior [Lipsch-Wijnen and Dirkx 2022], a análise da perspectiva do estudante sobre o feedback [Mandouit and Hattie 2023], classificação automática de mensagens de feedback educacional [Osakwe et al. 2022, Batista et al. 2022, Cavalcanti et al. 2021] e em articulação com outros modelos, como o de Nicol e Macfarlane-Dick (2006), para realizar uma análise da qualidade do feedback educacional [Cavalcanti et al. 2023].

No contexto da educação em programação, o modelo de Hattie e Timperley tem sido empregado tanto como base conceitual para a definição de feedback quanto como estrutura para classificar suas funções pedagógicas. [Ott et al. 2016] discutem a adaptação de princípios de feedback efetivo ao contexto de programação introdutória, enquanto [Keuning et al. 2018] utilizam os níveis de tarefa, processo e autorregulação para analisar

o potencial do feedback automatizado em exercícios de programação.

Estudos recentes indicam que LLMs têm sido explorados para gerar, avaliar e personalizar feedback em atividades de programação, tanto em abordagens automatizadas quanto híbridas com participação docente. Trabalhos como [Tang et al. 2024] e [Phung et al. 2025] destacam o potencial da combinação entre IA e intervenção humana para ampliar a qualidade, escalabilidade e confiabilidade do feedback. Já [Barros et al. 2025] analisam estratégias de *prompting* e parâmetros de geração com o modelo Llama 3, enquanto [Nguyen and Allan 2024] mostram que o GPT-4 pode produzir feedback útil e individualizado, embora ainda apresente limitações. Em conjunto, esses estudos reforçam o potencial dos LLMs como apoio ao feedback formativo em programação, desde que acompanhados por validação pedagógica e mediação docente [Tang et al. 2024, Phung et al. 2025, Barros et al. 2025, Nguyen and Allan 2024].

Além disso, trabalhos com LLMs passaram a investigar a capacidade dos modelos em produzir feedback formativo específico para programação. [Mueller et al. 2025] exploraram um tutor de programação baseado em LLM, com *prompts* alinhados aos modelos de Hattie e Timperley. O estudo apresentou uma preferência dos estudantes por feedback gerado por LLM contextualizado a partir do histórico de interações. Por sua vez, [Lohr et al. 2024] investigaram se LLMs conseguem gerar tipos específicos de feedback para tarefas introdutórias de programação, com base em taxonomias de feedback definidas por [Keuning et al. 2018]. O estudo destaca que LLMs podem ampliar as possibilidades de feedback rico e individualizado, mas também demandam controle pedagógico sobre o tipo, o conteúdo e a adequação da resposta gerada [Lohr et al. 2024].

Dessa forma, os trabalhos relacionados evidenciam que o feedback formativo deve ir além da simples indicação de erros e acertos, assumindo papel central na orientação da aprendizagem, na autorregulação e na definição de próximos passos para o estudante. Modelos como os de [Shute 2008], [Nicol and Macfarlane-Dick 2006] e, especialmente, [Hattie and Timperley 2007], oferecem bases conceituais importantes para estruturar feedbacks claros, específicos e pedagogicamente orientados. No contexto da programação, estudos recentes indicam que LLMs apresentam potencial para ampliar a oferta de feedback individualizado, escalável e contextualizado, embora ainda demandem validação pedagógica, controle sobre a qualidade das respostas e mediação docente.

2.1. Pergunta de Pesquisa

Esta pesquisa tem como objetivo comparar as percepções dos estudantes em relação à qualidade, clareza, utilidade e potencial formativo de dois tipos de feedback formativo aplicados a uma atividade avaliativa de POO. O primeiro feedback foi elaborado manualmente pelo professor da disciplina, e o segundo feedback foi gerado de forma automatizada por uma LLM com base no modelo teórico de Hattie e Timperley (2007). Assim, a pergunta de pesquisa levantada no trabalho é:

“Como o feedback formativo gerado por uma LLM, com base no modelo de Hattie e Timperley, é percebido pelos estudantes em comparação com o feedback elaborado pelo professor em uma atividade avaliativa de Programação Orientada a Objetos?”

3. Metodologia

A Figura 1 apresenta uma visão geral do procedimento metodológico adotado no estudo, sintetizando as principais etapas do estudo. Inicialmente, os estudantes realizaram uma atividade avaliativa de POO, que primeiramente foi corrigida manualmente pelo professor e gerado o feedback manual. Em seguida, as submissões dos estudantes foram anonimizadas e utilizadas como entrada para o LLM, juntamente com o enunciado da atividade e o *prompt* apresentado no Quadro 1.

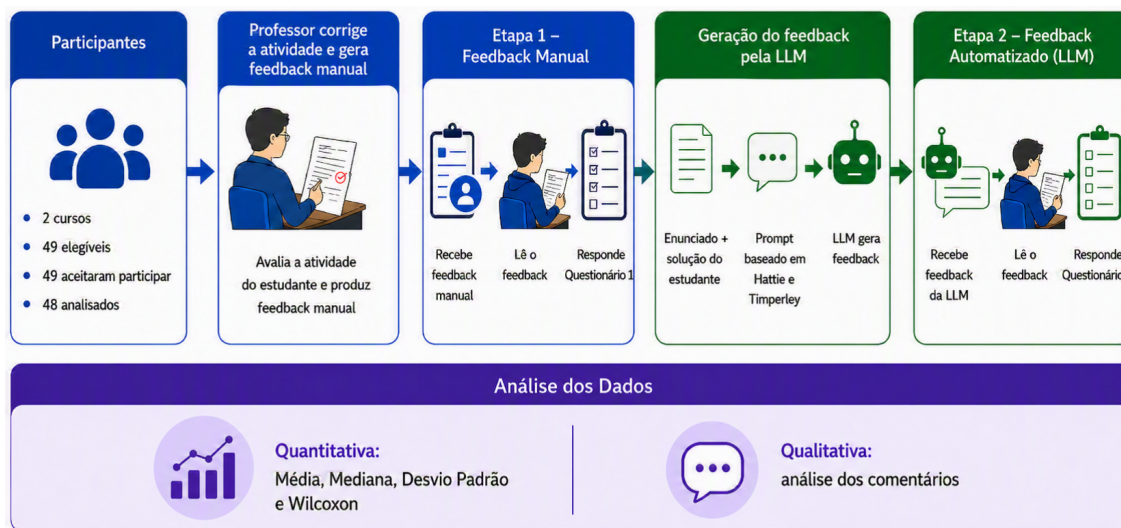


Figura 1. Fases do método de pesquisa.

O estudo foi conduzido em duas etapas presenciais com cada turma participante. Na primeira etapa, os estudantes receberam o feedback elaborado manualmente pelo professor e, em seguida, responderam ao instrumento de percepção. Posteriormente, foi disponibilizado o feedback gerado pela LLM, seguido da aplicação do mesmo instrumento, possibilitando a comparação das percepções atribuídas aos dois tipos de feedback. Durante todo o experimento, os participantes não foram informados sobre a origem dos feedbacks recebidos, de modo a reduzir possíveis vieses em suas avaliações. Por fim, os dados coletados foram analisados quantitativamente, com base nos constructos definidos no instrumento, e qualitativamente, a partir dos comentários abertos dos participantes.

Para a geração do feedback automatizado, utilizou-se o ChatGPT, modelo GPT-5.5, acessado por meio do site oficial da OpenAI em maio de 2026. A ferramenta foi escolhida por ser frequentemente utilizada na literatura [Jacobsen and Weber 2025, Mueller et al. 2025, Nguyen and Allan 2024, Phung et al. 2025]. O ChatGPT foi utilizado por meio de sua interface web, sem alterações de quaisquer parâmetros, buscando representar um cenário de uso semelhante ao que seria adotado por professores em contextos reais de apoio à avaliação. O LLM recebeu como entrada o enunciado da atividade, a solução submetida pelo estudante e o *prompt* desenvolvido.

A elaboração do *prompt* utilizado neste estudo foi fundamentada em práticas de *prompt engineering* recomendadas pela OpenAI¹ e em estudos recentes que orientam a interação com LLMs por meio da definição explícita de papel, contexto, tarefa, estrutura

¹<https://developers.openai.com/api/docs/guides/prompt-engineering>

Quadro 1. Estrutura do *prompt* utilizado.

< *context* > Você é um professor especialista em Programação Orientada a Objetos com experiência em avaliação formativa. Sua função é corrigir atividades de programação aplicando rigorosamente o modelo de feedback de Hattie e Timperley (2007), estruturado em três perguntas orientadoras. < /*context* >

< *feedback_model* > O modelo de Hattie e Timperley organiza o feedback em torno de três perguntas:

Feed Up → “Onde devo chegar?”(Esclarece quais são os objetivos de aprendizagem e os critérios de sucesso esperados.)

Feed Back → “Como estou indo em relação ao objetivo?”(Análise do desempenho. Fornece informações específicas sobre o progresso do estudante, indicando acertos, erros ou estratégias equivocadas.)

Feed Forward → “Quais são os próximos passos?”(Orienta o estudante sobre os próximos passos, novas estratégias, aprofundamentos ou desafios que podem promover avanços contínuos)

< /*feedback_model* >

< *description* > O enunciado da atividade está em anexo. < /*description* >

< *task* > Ao receber uma atividade de POO, siga EXATAMENTE esta estrutura de resposta:

- **Feed Up — Objetivos da Atividade** - Liste os objetivos de aprendizagem que esta atividade deveria alcançar (ex: encapsulamento, herança, polimorfismo, abstração, coesão, etc.). Se o enunciado não estiver claro, derive os objetivos a partir do código entregue.
- **Feed Back — Análise do Desempenho** - O que está correto (cite trechos ou comportamentos específicos do código) - O que está incorreto ou ausente (cite linha/bloco e explique o problema técnico) - Conformidade com os requisitos do enunciado - Estratégias de design adotadas - Avalie a legibilidade, nomenclatura e organização do código
- **Feed Forward — Próximos Passos** - Forneça de 1 a 2 orientações concretas e priorizadas para evolução, no formato: [Prioridade Alta/Média/Baixa] Ação específica → Justificativa técnica → Exemplo mínimo de como implementar
- **Síntese Avaliativa** Apresente uma síntese do comentário gerado em três parágrafos, onde cada parágrafo irá apresentar uma parte do comentário anterior. 1) Apresente o Feed Up em parágrafo curto. Destaque os principais conceitos da atividades.; 2) Apresente um resumo do Feed Back gerado, apresentando os principais pontos positivos e, para cada erro encontrado, explicar de forma sucinta o erro e a classe em que o problema foi encontrado. 3) Apresente o Feed Forward de forma sucinta, destacando as orientações para melhoria da solução apresentada. Essa síntese deve ser escrita um texto sucinto, claro, objetivo, respeitoso e com uma linguagem não genérica. < /*task* >

< *constraints* > - NUNCA emita julgamentos sobre o estudante como pessoa; foque exclusivamente no trabalho;

- SEMPRE cite trechos ou elementos específicos do código ao fazer qualquer afirmação;
- Se o enunciado da atividade não for fornecido, solicite-o antes de iniciar a correção;
- Se a linguagem de programação não for identificável, pergunte antes de prosseguir;
- Mantenha tom acadêmico, direto e construtivo em toda a resposta;
- Exemplos de código no *Feed Forward* DEVEM ser funcionais e minimais;
- Caso estudante apresente arquivos vazios ou partes do código faltantes, informe no feedback;

< /*constraints* >

da resposta e restrições pedagógicas [Alyahyan et al. 2025, Jacobsen and Weber 2025, Lohr et al. 2024, Yousef et al. 2025, Mueller et al. 2025]. Além disso, foram incorporados princípios do modelo de feedback de Hattie e Timperley (2007), com o objetivo de orientar a geração de feedback formativo alinhado às funções pedagógicas propostas pelos autores.

3.1. Participantes

O estudo foi conduzido durante a disciplina de POO, ofertada a estudantes de dois cursos superiores, Bacharelado em Ciência da Computação ($n = 22$) e Tecnologia em Sistemas para Internet ($n = 27$), com faixa etária entre 18 e 40 anos, com mediana de $M_d = 20$ anos. Ambas as disciplinas possuem o mesmo professor responsável e são ministradas na mesma instituição de ensino superior. A atividade avaliativa utilizada no estudo consiste em uma prova regular da disciplina que envolve conceitos centrais da disciplina, como abstração, classes, atributos, métodos, encapsulamento e herança. Foram considerados elegíveis ao estudo os estudantes que realizaram a atividade avaliativa ($n = 49$), aceitaram voluntariamente participar do estudo ($n = 49$) e responderam aos dois questionários de percepção aplicados ($n = 48$).

3.2. Instrumento de coleta de dados

O instrumento para a coleta da percepção do feedback utilizado neste estudo foi desenvolvido com base na literatura consolidada sobre feedback educacional e autorregulação da aprendizagem [Carless and Boud 2018, Nicol and Macfarlane-Dick 2006], conforme apresentado em um trabalho recente [Xavier et al. 2026]. O instrumento é composto por 14 itens respondidos em uma escala Likert de sete pontos, na qual valores menores indicam menor concordância e valores maiores, maior concordância. Ao final, foi incluída uma questão aberta para que os participantes registrassem suas percepções, impressões e sugestões sobre o feedback recebido.

Para este estudo, os constructos foram adaptados ao contexto do feedback formativo gerado por LLM em atividades de programação, tendo como principal referência o modelo de Hattie e Timperley. A adaptação preservou a base conceitual do instrumento original, mas reorganizou os itens para alinhá-los às funções pedagógicas do feedback descritas por Hattie e Timperley. Os itens e constructos analisados são apresentados no Quadro 2 e descritos a seguir.

1. **Clareza do Feedback** - Avalia a qualidade da comunicação do feedback. Busca medir principalmente se ele é inteligível e bem dimensionado
2. **Conexão com Metas (*Feed Up*)** - Reflete a capacidade do feedback de orientar o aluno em relação a onde ir. Verifica se o estudante compreendeu os objetivos, critérios e metas da atividade.
3. **Orientação para ação (*Feed Forward*)** - Avalia se o feedback apresenta estratégia e ações para melhoria da solução apresentada.
4. **Motivação e Autoeficácia** - Avalia os efeitos afetivos do feedback, como motivação, autoconfiança e crença na capacidade de executar as estratégias sugeridas
5. **Monitoramento do Progresso (*Feed Back*)** - Avalia se o estudante compreende seu desempenho atual em relação à meta
6. **Qualidade Relacional Professor-Aluno** - Avalia a dimensão sociointeracional do feedback, incluindo diálogo, reconhecimento individual e percepção de cuidado docente.

O coeficiente ômega de McDonald foi utilizado para avaliar a consistência interna dos constructos analisados, por ser uma medida de confiabilidade baseada em modelos fatoriais e menos restritiva que o alfa de Cronbach quanto à suposição de tau-equivalência

Quadro 2. Questão para avaliação da percepção do estudante sobre os feedbacks recebidos.

Constructo	Questão
Clareza do Feedback	1 - O feedback que recebi é facilmente compreensível. 2 - O feedback que recebi está em um tamanho adequado.
Conexão com Metas (<i>Feed Up</i>)	3 - Posso conectar o feedback que recebi com as metas desejadas (padrões) para minhas atividades da disciplina. 4 - O feedback que recebi me ajuda a entender melhor as atividades da disciplina. 7 - O feedback que recebi me ajuda a ajustar minhas metas para as atividades da disciplina.
Orientação para Ação (<i>Feed Forward</i>)	8 - O feedback que recebi me ajuda a desenvolver e ajustar minhas estratégias de aprendizagem. 9 - O feedback que recebi dá conselhos sobre o que posso fazer para atingir o desempenho desejado.
Motivação e Autoeficácia	5 O feedback que recebi me motiva a trabalhar em direção a uma meta desejada. 6 - O feedback que recebi ajuda a construir minha autoconfiança. 10 - Sinto-me confiante em executar as estratégias de aprendizagem sugeridas pelo feedback.
Monitoramento do Progresso - (<i>Feed Back</i>)	11 - O feedback que recebi me mostra meu progresso atual.
Qualidade Relacional Professor-Aluno	12 - O feedback que recebi incentiva o diálogo entre eu e o professor. 13 - O feedback que recebi mostra que meu professor entende meus pontos fortes e fracos. 14 - O feedback que recebi me faz sentir que meu professor se importa comigo.

entre os itens. Esse coeficiente é uma alternativa mais adequada para instrumentos compostos por itens com diferentes cargas fatoriais [Dunn et al. 2014]. Como critério interpretativo, adotou-se que valores de ω iguais ou superiores a 0,70 indicam confiabilidade aceitável, valores acima de 0,80 indicam boa confiabilidade e valores superiores a 0,90 indicam confiabilidade elevada, embora possam sugerir possível redundância entre itens.

Tabela 1. Resultado do ω de McDonald para os constructos utilizados.

Constructo	ω McDonald (Prof)	ω McDonald (LLM)
Clareza	0,8021	0,8259
Conexão com Metas	0,8179	0,9011
Orientação para Ação	0,8877	0,8031
Motivação/Autoeficácia	0,8347	0,8850
Monitoramento do Progresso	N/A	N/A
Qualidade Relacional	0,7951	0,9005

A análise dos dados foi realizada de duas formas complementares. A primeira foi uma análise quantitativa das respostas ao questionário, com o objetivo de comparar os constructos avaliados. A segunda foi uma análise qualitativa por meio da leitura e da análise do conteúdo dos comentários abertos, buscando compreender com maior profundidade as percepções, justificativas e sugestões apresentadas pelos estudantes.

Para analisar a percepção dos estudantes em relação aos constructos avaliados no instrumento, foram calculadas inicialmente medidas descritivas para cada dimensão, incluindo média simples, mediana e desvio padrão. A média foi utilizada para sintetizar o escore geral atribuído pelos participantes a cada constructo, enquanto a mediana permitiu observar a tendência central das respostas de forma menos sensível a valores extremos. O desvio padrão, por sua vez, foi empregado para verificar o grau de dispersão das respostas, indicando se houve maior ou menor concordância entre os estudantes.

Por fim, aplicou-se o teste não paramétrico de Wilcoxon para amostras pareadas. Esse teste foi escolhido por ser adequado para comparar medidas relacionadas, especialmente quando os dados provêm de escalas do tipo *Likert* ou quando não se assume a normalidade da distribuição das diferenças entre as condições analisadas [Sidney 1957]. A magnitude da diferença entre as condições pareadas foi estimada pelo método de Hodges-Lehmann (ΔHL), calculado como a pseudomediana das médias de Walsh das diferenças pareadas. O intervalo de confiança de 95% foi obtido pelo método exato, por inversão do teste de Wilcoxon, sem utilização de *bootstrap* [Hodges and Lehmann 1963]. Valores positivos indicam maior percepção favorável ao feedback gerado pela LLM, enquanto valores negativos indicam maior percepção favorável ao feedback do professor.

3.3. Cuidados Éticos

Este estudo foi aprovado pelo Comitê de Ética em Pesquisa (CEP), sob o número 85867624.1.0000.9547, e envolveu estudantes regularmente matriculados nas disciplinas de POO dos cursos de Bacharelado em Ciência da Computação e Tecnologia em Sistemas para Internet, cuja participação foi voluntária. Antes da geração dos *feedbacks* pela LLM, as submissões dos estudantes foram anonimizadas, de modo a remover informações que pudessem identificar os participantes.

4. Resultados

A Tabela 2 apresenta os resultados de cada métrica calculada. Os resultados mostram que, pelo teste de Wilcoxon, na maioria dos constructos avaliados, não há evidências suficientes para afirmar que os participantes perceberam diferenças significativas entre o feedback do professor e o feedback gerado pela LLM. Para Clareza, Conexão com Metas, Motivação/autoeficácia, Monitoramento do Progresso e Qualidade Relacional, os *p-value* foram superiores a 0,05, indicando ausência de evidência estatística suficiente para afirmar diferenças robustas entre as duas condições. Além disso, os intervalos de confiança da estimativa de Hodges-Lehmann incluem o valor zero, o que reforça a interpretação de que as diferenças observadas nesses constructos são pequenas e incertas quanto à direção.

A única exceção foi em relação ao constructo Orientação para Ação, que apresentou diferença estatisticamente significativa, com $W = 130,0$ e $p = 0,011$. Como esse constructo está relacionado à capacidade do feedback de oferecer indicações práticas sobre como melhorar a solução submetida, quais aspectos revisar e quais próximos passos seguir, o resultado pode indicar que o feedback gerado por LLM foi percebido como mais orientado à ação, um constructo relacionado ao *Feed Forward* do modelo de Hattie e Timperley. Para este constructo, a estimativa de Hodges-Lehmann foi de 1.00, indicando que o feedback da LLM foi avaliado, em termos típicos, cerca de 1.0 ponto acima do feedback do professor. O intervalo de confiança de 95% variou de 0.250 a 1.50, sugerindo que a diferença estimada pode variar de 0.250 até uma vantagem de 1.50 ponto para o LLM.

Tabela 2. Resultados dos dados dos constructos calculados.

Constructo	Média / Mediana (Desvio Padrão) (Prof)	Média / Mediana (Desvio Padrão) (LLM)	Wilcoxon	p_value	ΔH_L [IC 95%]
Clareza	6.11 / 6.250 (0.827)	6.01 / 6.500 (1.054)	302.5	0.647	0.000 [-0.500 , 0.250]
Conexão com Metas	6.06 / 6.333 (0.864)	6.12 / 6.333 (0.945)	318.0	0,822	0.000 [-0.333 , 0.500]
Orientação para Ação	5.53 / 5.500 (1.290)	6.13 / 6.000 (1.016)	130.0	0.011	1.000 [0.250, 1.500]
Motivação/Autoeficácia	5.53 / 6.000 (1.204)	5.81 / 6.000 (0.905)	238.0	0.139	0.333 [-0.167, 0.833]
Monitoramento do Progresso	6.02 / 7.000 (1.263)	5.83 / 6.000 (1.310)	158.0	0,470	0.000 [-1.000, 0.500]
Qualidade Relacional	5.74 / 5.667 (1.122)	5.72 / 6.000 (1.137)	354.5	0.830	0.000 [-0.500 , 0.500]
Média Geral	5,83	5,94	-	-	

5. Discussão

Os resultados quantitativos convergem com [Xavier et al. 2026], que também não observaram diferenças significativas na percepção dos estudantes entre feedbacks produzidos com e sem o apoio de LLM em diferentes constructos, como: conhecimento, motivação, metas, estratégias, produtos, monitoramento e relação com o professor. Contudo, os resultados divergem parcialmente de [Er et al. 2025], que identificaram maior utilidade percebida do feedback do professor em comparação ao feedback gerado por IA, além de um melhor desempenho no grupo que recebeu feedback docente.

É importante destacar que a ausência de significância estatística não indica, necessariamente, equivalência entre os feedbacks, mas apenas que os dados analisados não apresentaram evidências suficientes de diferenças estatisticamente robustas. Assim, os resultados quantitativos sugerem que os estudantes perceberam os dois tipos de feedback de forma semelhante na maioria das dimensões avaliadas.

Na análise qualitativa, os comentários indicam que os estudantes avaliaram positivamente ambos os tipos de feedback, mas perceberam diferenças entre eles. O feedback do professor foi descrito como mais claro, direto e coerente com o desempenho na avaliação, ajudando a identificar erros, pontos fortes e dificuldades específicas. Essa percepção aparece em comentários como P27: “o feedback fornecido pelo professor tornou mais claro meus pontos fortes e meus pontos fracos na construção do código” e P24: “foi bem direto e ao ponto do que devo melhorar”. Além disso, alguns estudantes destacaram que o comentário do professor demonstrava preocupação com a aprendizagem e contribuía para organizar os estudos.

Apesar da avaliação positiva, os comentários indicam que o feedback do professor P26: “poderia ser mais detalhado” e incluir orientações mais práticas para a correção dos erros. Alguns estudantes afirmaram que o retorno deveria P9: “deixar um pouco mais claro onde foi o erro e por que está errado”. Outro sugeriu incluir uma explicação prática

de como corrigir os erros do código, P35: *“como uma miniaula resumida explicando o problema de forma mais clara”*. Assim, embora valorizado pela objetividade, o retorno docente ainda poderia oferecer maior orientação sobre os próximos passos.

Em relação ao feedback produzido pela LLM, os estudantes destacaram o maior nível de detalhamento e a presença de orientações para melhorar o código. Comentários como P1: *“detalhou bem mais o que precisa ser corrigido”*, P24: *“ficou mais exemplificado do que eu errei”* e P26: *“mostrou o que preciso consertar e como devo fazer isso”* indicam que esse retorno foi percebido como mais explicativo, prático e voltado para a correção dos problemas. Para alguns estudantes, o feedback da LLM teve maior contribuição para compreensão dos erros e alternativas para melhorar a solução.

No entanto, o feedback da LLM também recebeu críticas importantes. A principal limitação apontada foi o excesso de texto, o que tornou a leitura cansativa ou menos objetiva para alguns alunos. Um estudante afirmou que P8: *“Apesar de ser detalhado e explicativo, é exagerado, (...), quando estava no final do feedback já não lembrava muito bem do início.”*. Outro comentário reforça P17: *“o maior problema desse feedback é o excesso de texto, se a resposta fosse menor e mais direta ao ponto seria melhor”*. Além disso, alguns estudantes consideraram que o texto é P38: *“muito técnico”* e P31: *“muito genérico”* e que P41: *“faltou um pouco de didática e algo mais pessoal”*, o que indica que o detalhamento, quando mal organizado, pode prejudicar a clareza e o engajamento.

De modo geral, os comentários sugerem que ambos os formatos apresentam contribuições complementares. O feedback do professor foi mais valorizado pela clareza, objetividade, coerência com a avaliação e proximidade com o aluno, enquanto o feedback da LLM foi reconhecido pelo melhor detalhamento, explicação dos erros e sugestões de melhoria. Entretanto, os estudantes indicam que o feedback ideal estaria em equilíbrio entre os dois: suficientemente detalhado para orientar a aprendizagem, mas organizado, objetivo, personalizado e com linguagem acessível. Essa conclusão aparece em comentário sobre o feedback da LLM: P38: *“achei bem completo, porém extremamente técnico. Acredito que o equilíbrio entre o primeiro feedback e esse seria o ideal”*.

O resultado da análise qualitativa se alinha às conclusões de [Yu and Xie 2025], que observaram que o ChatGPT produziu feedback mais detalhado e com mais explicações, mas também identificaram limitações relacionadas à linguagem, à extensão e à adequação ao nível dos estudantes. Também se alinha com [Lohr et al. 2024], ao indicar que LLMs apresentam potencial para gerar feedback formativo em programação, mas podem produzir informações excessivas ou até pouco adequadas para estudantes iniciantes. Assim, a principal contribuição dos resultados não está em afirmar a substituição do professor pela LLM, mas em evidenciar uma complementaridade entre os dois tipos de feedback, alinhada com a conclusão de [Er et al. 2025], que defendem modelos híbridos de feedback, nos quais a IA pode oferecer uma primeira versão ou apoiar aspectos mais operacionais, enquanto o professor revisa, personaliza e assegura a adequação pedagógica.

Por fim, esse resultado aponta para uma articulação importante entre os dados quantitativos e as evidências qualitativas. O constructo Orientação para Ação foi o único que apresentou diferença estatisticamente significativa, indicando que os estudantes perceberam o feedback da LLM como mais útil para indicar caminhos de correção e melhoria da atividade. Essa percepção também apareceu nos comentários abertos, nos quais

os participantes destacaram o maior detalhamento, a clareza das explicações e a presença de orientações sobre o que deveria ser corrigido. Entretanto, esse mesmo detalhamento também foi apontado como uma limitação, uma vez que o feedback da LLM foi percebido por alguns estudantes como mais longo e menos objetivo. Assim, o desafio está em equilibrar a quantidade de informação e oferecer orientações suficientemente detalhadas, mantendo o retorno claro, direto e adequado às necessidades dos estudantes.

5.1. Ameaça à validade

Ameaças à validade devem ser consideradas na interpretação dos resultados. Quanto à validade interna, destaca-se o possível efeito de ordem, pois todos os estudantes receberam primeiro o feedback do professor e, depois, o feedback gerado pela LLM, o que pode ter influenciado a avaliação do segundo feedback. Em relação à validade externa, o estudo foi realizado em uma única disciplina, em um contexto institucional específico e com uma atividade de POO, limitando a generalização dos achados para outros cenários educacionais. Na validade de conclusão, o tamanho da amostra pode ter reduzido o poder estatístico das análises, fazendo com que a ausência de significância em alguns construtos deva ser interpretada com cautela. Por fim, quanto à validade de construto, embora o questionário tenha sido adaptado de estudo anterior e apresentado consistência interna adequada, novas aplicações e análises adicionais são necessárias para fortalecer sua validade no contexto de atividades de programação.

6. Considerações Finais

Este estudo comparou a percepção dos estudantes sobre o feedback formativo gerado por uma LLM e pelo professor em atividades de programação. Os resultados quantitativos não revelaram diferenças estatisticamente significativas entre os dois tipos de feedback. Na análise qualitativa, o feedback do professor foi valorizado pela objetividade e adequação ao contexto da disciplina, enquanto o da LLM foi considerado mais detalhado, explicativo e estruturado, sobretudo na identificação de erros e na orientação para melhorias.

Esses resultados indicam que as LLMs apresentam potencial para apoiar práticas de avaliação formativa em programação, especialmente quando utilizadas como ferramenta complementar ao trabalho docente. Assim, os achados reforçam que seu uso não deve ser compreendido como substituição à atuação do professor, mas como uma possibilidade de ampliar a oferta de feedback individualizado, detalhado e oportuno. Em disciplinas de programação, nas quais a correção manual pode demandar elevado esforço docente, esse tipo de ferramenta pode contribuir para tornar o processo avaliativo mais escalável, preservando a importância da mediação pedagógica humana.

Como trabalhos futuros, pretende-se ampliar o número de participantes e aplicar o estudo comparativo em diferentes turmas, disciplinas e níveis de ensino, a fim de verificar a consistência dos resultados em outros contextos educacionais. Também se propõe investigar o impacto do feedback gerado por LLM no desempenho dos estudantes ao longo do tempo, considerando evidências de aprendizagem, revisão de código e melhoria em novas submissões, além de comparar diferentes modelos de LLM, estratégias de *prompting* e estruturas teóricas de feedback. Outra possibilidade é desenvolver e avaliar um sistema computacional que integre LLMs, rubricas avaliativas, análise automática de código e supervisão docente, permitindo ao professor revisar, ajustar e validar o feedback antes de disponibilizá-lo aos estudantes.

Agradecimentos

Os autores agradecem o apoio do Instituto Federal de Educação, Ciência e Tecnologia Goiano - IF Goiano e do Programa de Pós-graduação da Faculdade de Computação da Universidade Federal de Uberlândia (PPGCO/FACOM/UFU).

Declaração sobre uso de Inteligência Artificial

Os autores declaram que ferramentas de Inteligência Artificial generativa foram utilizadas como objeto de investigação deste estudo. O ChatGPT foi utilizado para a geração dos *feedbacks* formativos analisados pelos estudantes, conforme descrito na metodologia, geração de imagem e como apoio à escrita acadêmica e à revisão textual. A IA não é considerada autora ou coautora deste trabalho, cabendo aos autores toda a responsabilidade pelo conteúdo apresentado.

Referências

- Alyahyan, E., Bikanga Ada, M., and Lever, J. (2025). WIP: A Pedagogical Prompt Engineering Framework for LLM-Based Feedback in Higher Education (PPE-LLM). pages 1–5.
- Barros, J., Moraes, L. O., Oliveira, F., and Delgado, C. A. D. M. (2025). Large Language Models Generating Feedback for Students of Introductory Programming Courses. In Cristea, A. I., Walker, E., Lu, Y., Santos, O. C., and Isotani, S., editors, *Artificial Intelligence in Education*, pages 421–433, Cham. Springer Nature Switzerland.
- Batista, H. H. N., Cavalcanti, A. P., Miranda, P., Nascimento, A., and Mello, R. F. (2022). Classificação Multi-classe para Análise de Qualidade de Feedback. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1114–1125. SBC.
- Brooks, C., Huang, Y., Hattie, J., Carroll, A., and Burton, R. (2019). What Is My Next Step? School Students’ Perceptions of Feedback. *Frontiers in Education*, 4.
- Carless, D. and Boud, D. (2018). The development of student feedback literacy: enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43:1315–1325.
- Cavalcanti, A. P., Mello, R. F., Miranda, P., Nascimento, A., and Freitas, F. (2021). Utilização de Recursos Linguísticos para Classificação Automática de Mensagens de Feedback. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 861–872. SBC.
- Cavalcanti, A. P., Rolim, V. B., Gašević, D., and Mello, R. F. (2023). A Comparative Analysis Between Good Feedback Descriptors on Online Courses. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1512–1523. SBC.
- Dunn, T. J., Baguley, T., and Brunsden, V. (2014). From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3):399–412.
- Efan, E., Krismadinata, K., Jama, J., and Mulya, R. (2023). A Systematic Literature Review of Teaching and Learning on Object-Oriented Programming Course. *International Journal of Information and Education Technology*, 13:302–312.

- Er, E., Akçapınar, G., Bayazıt, A., Noroozi, O., and Banihashem, S. K. (2025). Assessing student perceptions and use of instructor versus AI-generated feedback. *British Journal of Educational Technology*, 56(3):1074–1091. _eprint: <https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13558>.
- Hattie, J. and Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, 77(1):81–112.
- Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34(2):598–611.
- Jacobsen, L. J. and Weber, K. E. (2025). The Promises and Pitfalls of Large Language Models as Feedback Providers: A Study of Prompt Engineering and the Quality of AI-Driven Feedback. *AI*, 6(2):35.
- Keuning, H., Jeurig, J., and Heeren, B. (2018). A Systematic Literature Review of Automated Feedback Generation for Programming Exercises. *ACM Transactions on Computing Education*, 19:1–43.
- Lipsch-Wijnen, I. and Dirkx, K. (2022). A case study of the use of the Hattie and Timperley feedback model on written feedback in thesis examination in higher education. *Cogent Education*, 9(1):2082089. _eprint: <https://doi.org/10.1080/2331186X.2022.2082089>.
- Lohr, D., Keuning, H., and Kiesler, N. (2024). You’re (Not) My Type – Can LLMs Generate Feedback of Specific Types for Introductory Programming Tasks?
- Mandouit, L. and Hattie, J. (2023). Revisiting “*The Power of Feedback*” from the perspective of the learner. *Learning and Instruction*, 84:101718.
- Mueller, M., List, C., and Kipp, M. (2025). The Power of Context: An LLM-based Programming Tutor with Focused and Proactive Feedback. In *Proceedings of the 6th European Conference on Software Engineering Education*, ECSEE ’25, pages 1–10, New York, NY, USA. Association for Computing Machinery.
- Nguyen, H. and Allan, V. (2024). Using gpt-4 to provide tiered, formative code feedback. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, SIGCSE 2024, page 958–964, New York, NY, USA. Association for Computing Machinery.
- 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.
- Osakwe, I., Chen, G., Whitelock-Wainwright, A., Gašević, D., Pinheiro Cavalcanti, A., and Ferreira Mello, R. (2022). Towards automated content analysis of educational feedback: A multi-language study. *Computers and Education: Artificial Intelligence*, 3:100059.
- Ott, C., Robins, A., and Shephard, K. (2016). Translating Principles of Effective Feedback for Students into the CS1 Context. *ACM Trans. Comput. Educ.*, 16(1):1:1–1:27.
- Phung, T., Choi, H., Wu, M., Brooks, C., Gulwani, S., and Singla, A. (2025). Closing the Loop: An Instructor-in-the-Loop AI Assistance System for Supporting Student Help-Seeking in Programming Education. arXiv:2510.14457 [cs].

- Shute, V. (2008). Focus on Formative Feedback. *Review of Educational Research*, 78:153–189.
- Sidney, S. (1957). NONPARAMETRIC STATISTICS FOR THE BEHAVIORAL SCIENCES:. In *The Journal of Nervous and Mental Disease*, volume 125, page 497. Journal Abbreviation: The Journal of Nervous and Mental Disease.
- Silva, F. G. and Aranha, E. H. S. (2025). LLMs na Educação em Programação: Estratégias para Avaliação e Feedback Formativo. In *Simpósio Brasileiro de Educação em Computação (EDUCOMP)*, pages 81–87. SBC.
- Tang, X., Wong, S., Huynh, M., He, Z., Yang, Y., and Chen, Y. (2024). SPHERE: Scaling Personalized Feedback in Programming Classrooms with Structured Review of LLM Outputs. arXiv:2410.16513 [cs].
- 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., and Mello, R. F. (2026). Human Teacher vs. LLM-Generated Feedback in Secondary Education: A Comparative Study on Student Perceptions. In Tammets, K., Sosnovsky, S., Ferreira Mello, R., Pishtari, G., and Nazaretsky, T., editors, *Two Decades of TEL. From Lessons Learnt to Challenges Ahead*, pages 534–548, Cham. Springer Nature Switzerland.
- Yousef, M., Mohamed, K., Medhat, W., Mohamed, E. H., Khoriba, G., and Arafa, T. (2025). BeGrading: large language models for enhanced feedback in programming education. *Neural Computing and Applications*, 37(2):1027–1040.
- Yu, H. and Xie, Q. (2025). Generative AI vs. teachers: Feedback quality, feedback uptake, and revision. *Language Teaching Research Quarterly*, 47:113–137.