AUTOGRADER COM IA: FEEDBACK AUTOMATIZADO EM AMBIENTES DE APRENDIZAGEM ATIVA

AI-powered Autograder: automated feedback in active learning environments.

ROMMEL VIEIRA CARNEIRO
UFMG

FREDERICO CESAR MAFRA PEREIRA
UFMG

RODRIGO BARONI DE CARVALHO
PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS (PUC MINAS)

ARTHUR CARVALHO RODRIGUES

Comunicação:
O XIV SINGEP foi realizado em conjunto com a 14th Conferência Internacional do CIK (CYRUS Institute of Knowledge) e com a SPARK (Society for Promoting Academic Research and Knowledge), em formato híbrido, com sede presencial na UNINOVE - Universidade Nove de Julho, no Brasil.

AUTOGRADER COM IA: FEEDBACK AUTOMATIZADO EM AMBIENTES DE APRENDIZAGEM ATIVA

Objetivo do estudo
Avaliar o efeito da aplicação de um sistema de feedback automatizado (autograder) com modelos de linguagem em atividades práticas de Engenharia de Software, considernado o envolvimento dos estudantes e a percepção de clareza e utilidade no processo formativo.

Relevância/originalidade
Turmas maiores e educação continuada pressionam a oferta de feedback individualizado. O estudo é original ao comparar, em sala de aula real, três cenários: sem correção automatizada, com correção por testes e com camada de LLM sobre árvore de critérios configurável.

Metodologia/abordagem
Relato de experiência com abordagem mista. Vinte estudantes tiveram commits acompanhados em quatro atividades distribuídas nos três cenários; onze responderam questionário Likert de quatro pontos. Análise descritiva das medidas e codificação temática das respostas abertas.

Principais resultados
A média de commits subiu de 4,55 para 6,90 e 6,00, mas a mediana permaneceu constante, indicando efeito concentrado nos extremos. A percepção evoluiu de forma consistente: suficiência de orientação e feed passou de 45% para 100% de concordância.

Contribuições teóricas/metodológicas
Articula a literatura de feedback formativo e educação por competências a um artefato computacional, propondo a CriteriaTree (estrutura hierárquica e configurável de critérios ponderados) e um desenho comparativo de três cenários replicável em outras disciplinas práticas.

Contribuições sociais/para a gestão
Oferece a gestores acadêmicos e corporativos um caminho de baixo custo para escalar o acompanhamento formativo sem ampliar proporcionalmente o corpo docente, com diretrizes de adoção gradual, curadoria pedagógica e uso ético de IA na formação tecnológica.

Palavras-chave: Inteligência Artificial, Feedback Automatizado, Educação Baseada em Competências, Metodologias Ativas, Autograder

AI-powered Autograder: automated feedback in active learning environments.

Study purpose
To assess the effect of an LLM-based automated feedback system (Autograder) on platform usage and on students' perceived clarity and usefulness of the feedback received in hands-on Software Engineering activities.

Relevance / originality
Growing class sizes and continuing education strain the delivery of individualized feedback. The study is original in combining automated test-based grading with LLM-generated feedback in a real classroom, comparing three usage scenarios.

Methodology / approach
Experience report with a mixed-methods approach. Twenty students had their commits tracked across four activities spanning the three scenarios; eleven answered a four-point Likert questionnaire. Descriptive analysis of measures and thematic coding of open-ended responses.

Main results
Mean commits rose from 4.55 to 6.90 and 6.00, but the median remained constant, indicating an effect concentrated in the tails. Perception improved consistently: sufficiency of guidance and feedback rose from 45% to 100% agreement.

Theoretical / methodological contributions
It proposes the CriteriaTree, a hierarchical structure of weighted criteria that makes automated grading configurable, and a replicable three-scenario comparative design. It distinguishes usage indicators from engagement measures, warning against conflating treatment and proxy.

Social / management contributions
It offers academic and corporate managers a low-cost architecture to scale formative follow-up on existing infrastructure, with guidelines for gradual adoption, pedagogical curation, and ethical use of AI in technology education.

Keywords: Artificial Intelligence, Automated Feedback, Competency-Based Education, Active Methodologies, Autograder

AUTOGRADER COM IA: FEEDBACK AUTOMATIZADO EM AMBIENTES DE APRENDIZAGEM ATIVA

1 Introdução

As transformações digitais têm redefinido as dinâmicas sociais, profissionais e educacionais, exigindo novas competências técnicas e comportamentais dos profissionais. Com isso, percebe-se um aumento de profissionais que procuram capacitação profissional tanto no ensino formal quanto no mercado de cursos livres com foco na evolução de carreira ou mesmo para a manutenção da empregabilidade.

Esse contexto traz novos desafios à formação de pessoas capazes de atuar com protagonismo em um cenário cada vez mais volátil, incerto e complexo. A área da educação vem sofrendo profundas mudanças com o fenômeno da convergência tecnológica e o processo de ensino e aprendizagem está se tornando mais centrado na experiência do ser humano (Mukul & Büyüközkan, 2023). As instituições e os educadores têm sido pressionados a reverem seus modelos pedagógicos, adequando-os a uma realidade na qual o aprendizado deve ser contínuo, flexível, personalizado e centrado no desenvolvimento de competências alinhadas com o mundo do trabalho (Levine & Patrick, 2019; Gervais, 2016). O volume cada vez maior e diverso de alunos, além do aumento na abordagem remota de ensino, traz novos desafios para a personalização e a eficácia do aprendizado. Essa realidade demanda metodologias ativas e um design instrucional que promova a formação de competências alinhadas às demandas do mercado de trabalho (Levine & Patrick, 2019).

O surgimento de assistentes de IA baseados em Grandes Modelos de Linguagem (LLMs) intensificou as oportunidades e os desafios na área da educação, exigindo novas estratégias de ensino, aprendizagem e avaliação (Raihan et al., 2025). Essas ferramentas podem apoiar professores no planejamento e geração de material didático, em pesquisas e tarefas de escrita, nas avaliações formativas e somativas, bem como apoiar aprendizes no processo de aprendizado criando questões para validação do conhecimento, explicando conceitos mais complexos e até assistindo em pesquisas e escrita de trabalhos (Kasneci et al., 2023). Por outro lado, o uso das IAs no processo de ensino e aprendizagem pode gerar grande dependência dos assistentes, problemas de qualidade nos *feedbacks*, limitações em tópicos avançados além de prejudicar a capacidade dos estudantes de resolver problemas de forma independente (Raihan et al., 2025).

O objetivo central deste trabalho é desenvolver mecanismos de correção automática de trabalhos em jornadas formativas que deem agilidade no processo de *feedback* aos alunos, promovendo maior envolvimento e uma experiência personalizada e de qualidade para um volume cada vez maior de participantes. Como objetivos específicos, busca-se investigar o impacto da aplicação destes mecanismos na efetividade do processo de ensino e aprendizagem, por meio da avaliação das percepções dos participantes quanto à clareza e utilidade dos *feedbacks* recebidos.

A pesquisa se justifica pela crescente necessidade de soluções educacionais que respondam com agilidade e qualidade às transformações digitais e às exigências do mercado de trabalho, especialmente em contextos que envolvem educação continuada e inserção profissional.

A partir dos objetivos delineados, formulam-se as seguintes hipóteses e perguntas de pesquisa, que orientam a condução deste estudo:

- H1: O *feedback* imediato aumenta o envolvimento do aluno nas atividades práticas.
- H2: O nível de conhecimento prévio influencia a necessidade de *feedback* automatizado.
- H3: O baixo interesse na disciplina reduz o impacto da aplicação do *feedback* automatizado.

Assim, busca-se responder às seguintes perguntas de pesquisa: (i) "O *feedback* automatizado aumenta o envolvimento dos alunos?"; (ii) "Melhora o desempenho e a clareza da aprendizagem?" e (iii) "Proporciona orientação estruturada e reduz o tempo de execução das tarefas?".

O artigo foi estruturado da seguinte maneira: após esta introdução, o Capítulo 2 apresenta a fundamentação teórica, discutindo os conceitos de *feedback* formativo, educação por competências e o papel das tecnologias digitais e da inteligência artificial no apoio à aprendizagem. O Capítulo 3 descreve a metodologia utilizada, detalhando o contexto da pesquisa, o delineamento do estudo, os instrumentos de coleta e os procedimentos de análise dos dados. Em seguida, O Capítulo 4 apresenta a estrutura do Autograder. O Capítulo 5 expõe os resultados e a análise, apresentando as evidências empíricas obtidas e sua interpretação à luz da literatura. Por fim, o Capítulo 6 reúne as conclusões e recomendações, destacando as contribuições do estudo, suas limitações e as perspectivas para pesquisas futuras sobre o uso de sistemas automatizados de *feedback* em contextos formativos.

2 Fundamentação Teórica

Neste capítulo, é apresentada uma revisão da literatura sobre a importância do *feedback* no processo de ensino e aprendizagem, as bases da educação por competências e as abordagens de metodologias ativas, assim como da aplicação de sistemas e Inteligência Artificial (IA) para a geração de *feedback* automatizado de atividades de avaliação formativa.

A Educação Baseada em Competências (EBC) pode ser entendida como uma abordagem de ensino voltada ao desenvolvimento de um conjunto de conhecimentos, habilidades, atitudes e valores associados a uma competência almejada (Gervais, 2016, p. 99). Os principais componentes da EBC estão relacionados com uma abordagem centrada no aprendiz, flexibilidade no caminho de aprendizagem, foco no desenvolvimento de habilidades, avaliação e *feedback* contínuos voltados para o desempenho do aprendiz (Gervais, 2016; Oroszi, 2020). A Educação 4.0 é caracterizada por várias características inovadoras: (1) educação independente de tempo e lugar para um aprendizado flexível; (2) aprendizado personalizado e baseado em projetos com foco na aplicação prática; e (3) avaliação contínua em vez de exames tradicionais (Mukul & Büyüközkan, 2023). A Aprendizagem Baseada em Projetos (PjBL) é um método pedagógico que une teoria e prática ao envolver os alunos em desafios do mundo real. No entanto, a adoção em larga escala da PjBL enfrenta desafios devido ao tempo e à expertise necessários para um *feedback* eficaz, especialmente com o aumento do número de alunos (Kusam, 2024).

O *feedback* é uma das ferramentas mais poderosas na conquista de resultados no processo de aprendizagem, porém, seu impacto pode ser positivo ou negativo, uma vez que o tipo de *feedback* e a forma como é fornecido afetam sua eficácia (Hattie & Timperley, 2007). O objetivo do *feedback* é reduzir a distância entre a compreensão e desempenho do aluno de uma meta estabelecida (Hattie & Timperley, 2007). Segundo Hattie e Timperley (2007), o *feedback* pode ser direcionado o que influencia sua eficácia, podendo estar em quatro níveis: nível da tarefa, nível do processo, nível de autorregulação e nível próprio. Conforme os autores, os melhores resultados de *feedback* acontecem ao se prover informações sobre uma tarefa e como melhorá-la de forma eficaz, ao passo que elogios, recompensas e punições foram menos eficazes. Em iniciativas baseadas em PjBL, *feedbacks* são essenciais para orientar os estudantes na realização das tarefas e ajudá-los a aprenderem a partir do processo (Kusam, 2024). No entanto, a natureza intensiva do *feedback* personalizado, que exige tempo e expertise específicos da área, é um grande desafio, especialmente em turmas grandes de programas de ciência da computação (Kusam, 2024).

Muitos trabalhos têm sido realizados para avaliar como simplificar o processo de *feedback* em iniciativas de ensino e aprendizagem por meio do uso de tecnologias, em especial, as novas inteligências artificiais generativas (Kasneci et al., 2023; Kusam, 2024; Chan et al., 2024; Raihan et al., 2025). Existem vantagens em criar soluções de *feedback* automatizado como escalabilidade do número de estudantes em iniciativas de ensino, melhoria na experiência de aprendizagem, redução no viés em avaliações; ao mesmo tempo que existem desafios como a criação de cultura de respostas menos preparadas por serem avaliadas de forma padronizada, além de incentivarem estudantes a apenas responderem perguntas ao invés de aprenderem conceitos (Guerra Hahn et al., 2021).

3 Metodologia

A pesquisa caracteriza-se como aplicada, de abordagem mista e natureza qualitativa-quantitativa (Creswell, 2014), configurando um relato de experiência. O estudo foi conduzido no âmbito de uma disciplina de Desenvolvimento de Interfaces Web, oferecida em um curso de Engenharia de Software, envolvendo quatro atividades práticas de programação com uso das linguagens HTML, CSS e JavaScript.

O experimento teve a participação de 20 estudantes, que tiveram suas entregas acompanhadas nas quatro atividades da disciplina. Ao final do processo, foi aplicado um questionário eletrônico, respondido por 11 estudantes, uma taxa de resposta de 55%. Todos os participantes registraram anuência ao Termo de Consentimento Livre e Esclarecido (TCLE) ao acessarem o questionário. Os dados foram coletados no período de agosto de 2025 a novembro de 2025.

As evidências foram obtidas em três cenários distintos: (i) Atividade sem Autograder: os estudantes realizaram a atividade sem qualquer forma de correção automatizada (TP3); (ii) Atividade com Autograder básico (sem LLM): a atividade contou com *feedback* automático baseado apenas em scripts de testes de software (TP4 e TP5); e (iii) Atividade com Autograder com LLM: a atividade incluiu avaliação por testes de software e *feedback* gerado por modelo de linguagem de larga escala (LLM), capaz de interpretar e comentar o código (TP6). Como o segundo cenário compreende duas atividades, os indicadores são apresentados por atividade e,

nas comparações dos cenários, os indicadores são agrupados por meio da média as atividades TP4 e TP5.

Foram coletados dois conjuntos de dados. O primeiro é formado por indicadores de interações dos alunos na plataforma do GitHub, como o número médio de *commits* por aluno em cada uma das atividades, sendo utilizado como um indicador de uso da ferramenta na realização da referida atividade. O segundo foi realizado por meio de questionário eletrônico com perguntas no modelo de escala *Likert* de quatro pontos, com opção explícita de "não consigo responder" para os casos em que o estudante não tenha vivenciado o cenário. Foram avaliadas as seguintes dimensões: (i) suficiência de orientação e feedback; (ii) clareza do processo, (iii) envolvimento na atividade; (iv) agilidade na entrega e (v) qualidade das entregas. Adicionalmente, foi incluída uma questão aberta utilizada para captar percepções qualitativas sobre a experiência no uso da ferramenta.

A análise quantitativa foi tratada por meio de estatística descritiva com apresentação de média, desvio-padrão, mediana e intervalo interquartil. Na análise qualitativas, as respostas "não consigo responder" foram tratadas como dados ausentes; os percentuais de concordância foram calculados sobre respostas válidas, com o total de participantes (n) de cada item explicitado nas tabelas. Nos itens comparáveis entre cenários, o número de casos completos nos três cenários é de nove.

A verificação do efeito sobre a carga docente e sobre a viabilidade de turmas maiores exige instrumentação do trabalho do professor e não foi contemplada neste trabalho, permanecendo como uma proposta futura de pesquisa.

4 Estrutura do Autograder

O sistema, denominado Autograder, foi projetado com objetivo de apoiar o trabalho do educador, aumentar o envolvimento dos estudantes, fomentando o trabalho contínuo em atividades práticas por meio de um *feedback* automatizado.

A função do sistema, proposto neste trabalho, é fornecer um processo de avaliação estruturado e um *feedback* formativo de forma instantânea. A arquitetura do sistema permite a configuração de rotinas de autocorreção flexíveis que se adaptam rigorosamente aos requisitos de uma atividade. A velocidade com que esse ciclo se completa, da submissão à entrega do *feedback*, é um diferencial, pois atua no momento de interesse do aluno e reduz a desmotivação causada pela espera por algum resultado.

Assim que o estudante realiza a entrega associada a uma atividade, o sistema é disparado para realizar a verificação do que foi solicitado para a atividade. Com base nos resultados da verificação, o Autograder gera um *feedback* formativo que informa e estimula o aluno a se manter no fluxo de aprendizagem, com o propósito de aumentar o envolvimento do aluno na atividade. Um diagrama esquemático é apresentado na Figura 1.

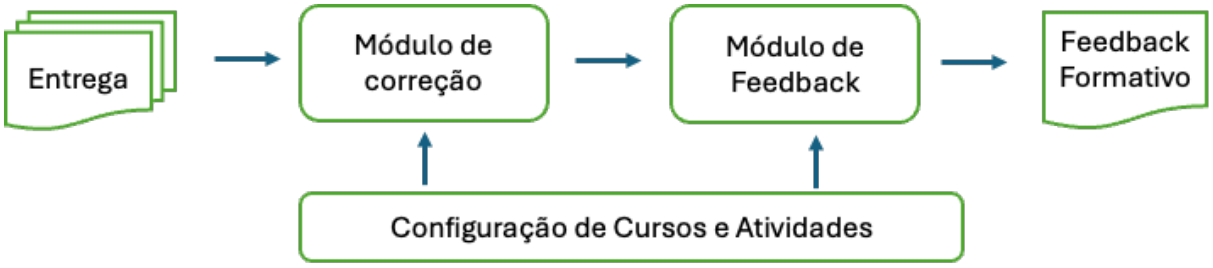


Figura 1. Arquitetura da solução do Autograder.

A solução do Autograder se baseia em um conjunto de scripts executados de forma integrada pela plataforma do GitHub, com atividades controladas por meio do GitHub Classroom, bem como no uso da *Application Programming Interface (API)* da Open AI para acesso aos modelos de IA. O diagrama com a descrição detalhada do fluxo de interação de educadores e estudantes com as plataformas envolvidas é apresentado na Figura 2.

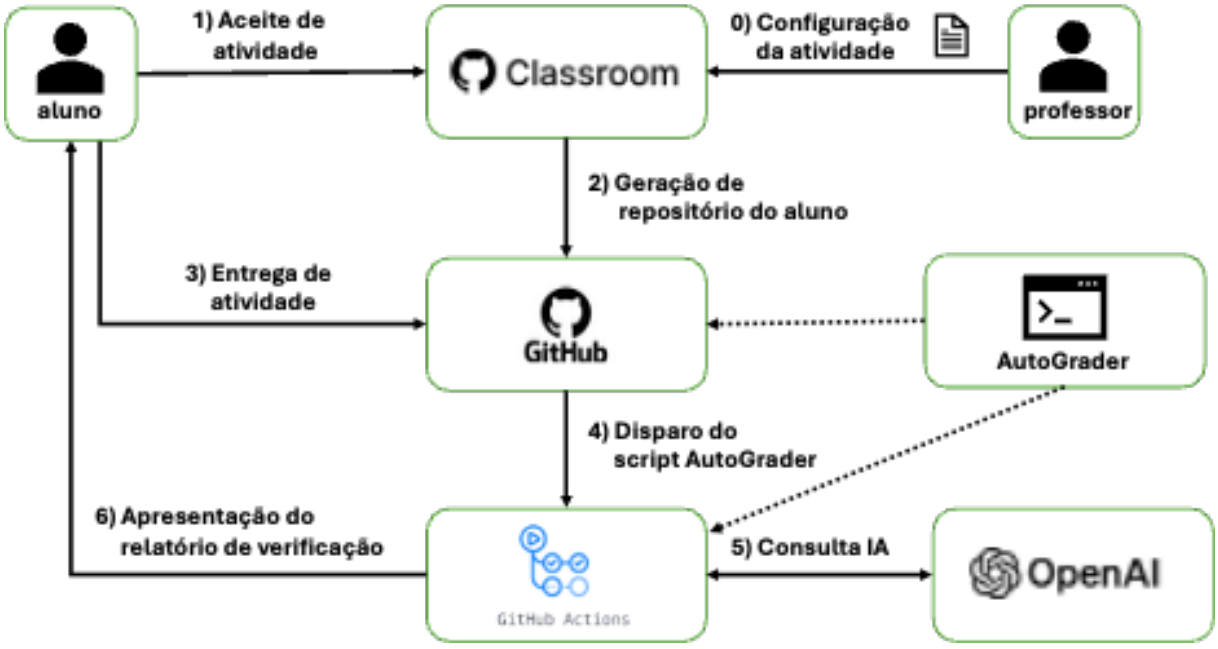


Figura 2. Fluxo de interação do Autograder

Para trabalhar com diferentes tipos de atividades, o Autograder possui um mecanismo de configuração flexível baseado em um conjunto de critérios que definem a lógica de correção e são organizados hierarquicamente por meio de uma estrutura de dados em árvore, a CriteriaTree. A árvore de critérios funciona como espinha dorsal para o processo de correção, de forma que todos os componentes do sistema se baseiam nas informações nela contidas.

O módulo de correção consulta a CriteriaTree para se orientar quanto aos testes a serem executados e sua respectiva ponderação na pontuação. Em seguida, o módulo de geração de *feedback* analisa a estrutura da árvore em conjunto com os resultados obtidos pelo aluno para construir o retorno adequado no contexto formativo.

A escolha de uma estrutura em árvore foi motivada pela necessidade de flexibilidade e pela capacidade de representar uma hierarquia de pontuação mais complexa com agrupamentos aninhados. A CriteriaTree, exemplificada na Figura 3, organiza os testes em grupos

denominados *Subjects*. Cada *Subject* possui um peso de pontuação atribuído (porcentagem da pontuação disponível em seu nível hierárquico) e pode conter outros *Subjects* aninhados. As folhas da árvore correspondem aos testes unitários executados sobre a submissão do aluno, cada um retornando uma pontuação.

Essa estrutura facilita a propagação da pontuação à medida que os testes nas folhas são executados, sua pontuação é agregada e ponderada recursivamente pelos *Subjects* pais, ascendendo na hierarquia até que o nó raiz consolide a nota final da submissão.

O nó raiz é subdividido em três nós fixos que estruturam o conjunto dos critérios de correção. São eles:

- Base - Contém testes para requisitos obrigatórios da atividade, correspondendo a 100% da pontuação;
- Bônus - Valoriza contribuições que, embora não obrigatórias, demonstram um envolvimento aprofundado e esforços que excedem o esperado;
- Penalidade - Identifica e subtrai pontos pela presença de elementos indesejados ou práticas inadequadas.

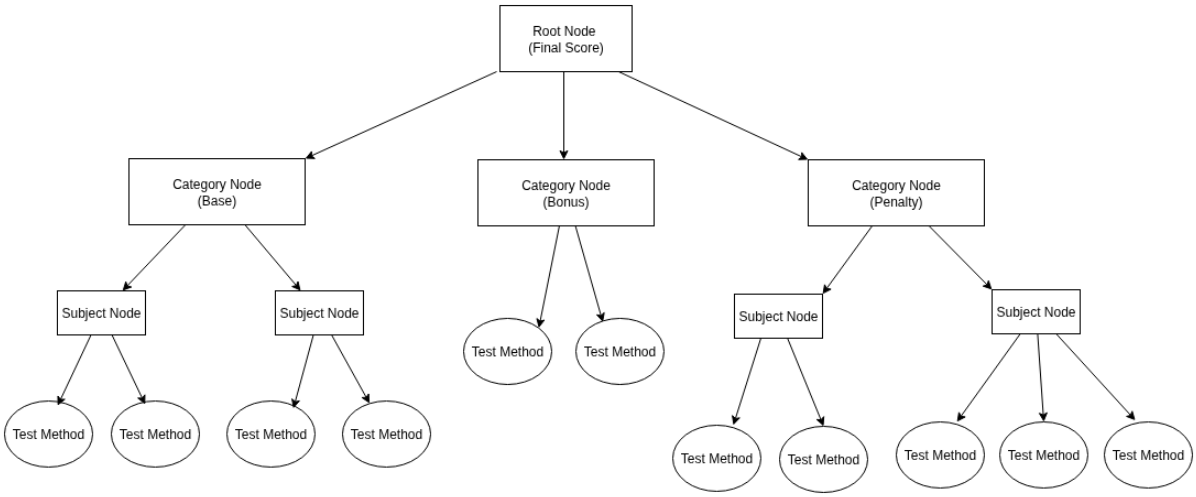


Figura 3. Árvore de Critérios (CriteriaTree)

O cálculo da nota final segue uma sequência lógica: primeiramente, a pontuação base é calculada. Se não atingir 100%, a pontuação de bônus é somada, limitando a nota final ao teto de 100%. Por fim, a pontuação de penalidade é subtraída do total acumulado com nota final mínima de 0. Entre esses nós de categoria e os testes, é permitida a criação de múltiplos níveis de *Subjects* aninhados. Ao final da execução da árvore em uma submissão, é possível ter uma visualização granular do aproveitamento do aluno, o que facilita a identificação de pontos de dificuldade.

Em uma arquitetura multiplataforma, os alunos transitam entre contextos teóricos de aprendizagem passiva (ministrados na plataforma LMS) e contextos práticos, como desafios e projetos (atividades). O Autograder se insere no contexto prático, servindo simultaneamente

como um reforço teórico por meio dos relatórios de avaliação. As atividades práticas são registradas no GitHub Classroom, onde repositórios de código são criados dinamicamente para cada aluno. O Autograder é configurado como uma ação (*action*) do GitHub Actions, uma ferramenta de automação nativa da plataforma GitHub. Essa automação é crucial, pois a submissão de código (*commit*) por um aluno a um repositório configurado desencadeia a execução imediata do fluxo de correção. O processo finaliza com a inserção de um arquivo de *feedback* no repositório do aluno e com a atualização de sua nota no ambiente do GitHub Classroom que oferece um painel consolidado para o educador.

Essa automação garante que o ciclo completo ocorra em questão de minutos, um fator determinante para manter o envolvimento do aluno no processo de aprendizado. Esse ciclo inclui a submissão de uma entrega para a atividade, a verificação por parte do módulo de correção e a geração e entrega do *feedback* por parte do módulo de *feedback*. A eficácia do Autograder é apoiada por sua integração nesse framework de aprendizagem mais amplo.

Na etapa final do processo, os resultados da execução da CriteriaTree são traduzidos em um texto formativo para o aluno. Esse texto é projetado para maximizar o impacto pedagógico, promovendo o envolvimento e a aprendizagem contínua. O sistema oferece duas modalidades de geração de *feedback*:

Feedback Padrão (sem IA) - Construído a partir dos dados brutos da correção. Este modo oferece um *feedback* estruturado, direto e objetivo. É ideal para cenários que exigem clareza e precisão imediatas sobre os erros e acertos. Um exemplo é apresentado na Figura 4.

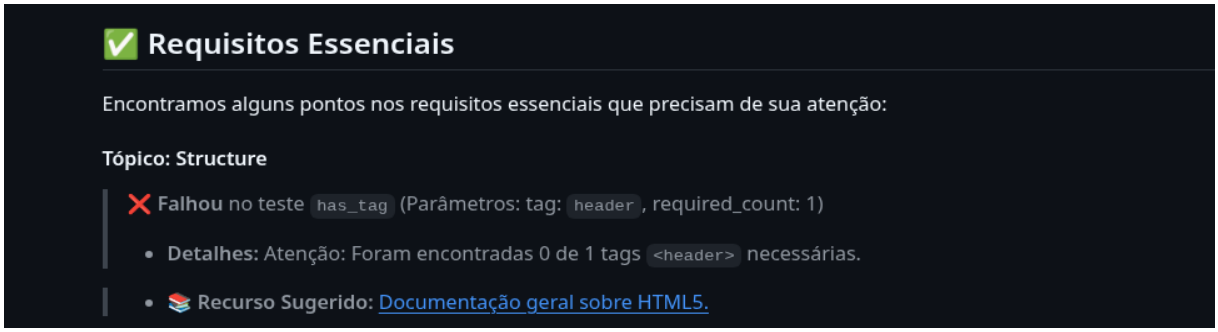


Figura 4. *Feedback* padrão (sem Inteligência Artificial)

Feedback por IA: Esta opção proporciona uma análise qualitativa e contextualizada, buscando interpretar relações de causalidade entre os erros e os conceitos subjacentes. Os resultados são apresentados em uma linguagem mais natural e com estrutura conversacional, o que pode aumentar a receptividade do aluno. Um exemplo é apresentado na Figura 5.

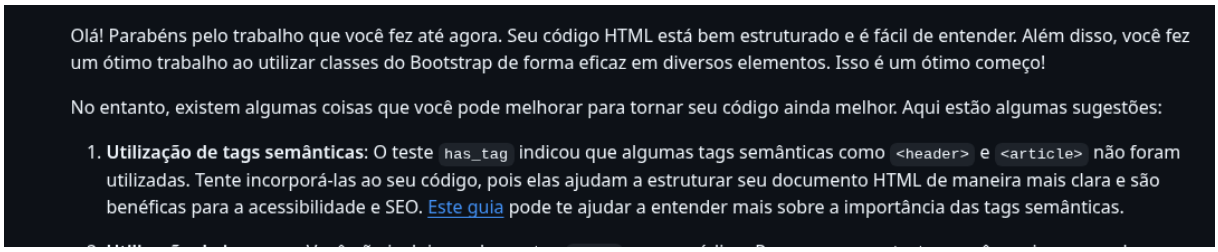


Figura 5. *Feedback* por Inteligência Artificial

A escolha entre os modos de *feedback* depende dos objetivos pedagógicos da atividade. A modalidade de *feedback* por IA se destaca por sua capacidade de personalização, permitindo a configuração de parâmetros como persona (ex: "tutor amigável"), tom de voz, nível de ajuda, contexto da atividade e outras orientações. Essa flexibilidade possibilita a geração de um *feedback* extremamente humanizado, assertivo e alinhado ao padrão de ensino desejado.

5 Resultados e Discussão

Esta seção apresenta os principais achados obtidos a partir da aplicação do sistema de *feedback* automatizado (Autograder) em três cenários distintos da disciplina de Desenvolvimento de Interfaces Web. Os resultados são organizados em três subseções: (i) indicadores de uso extraídos da plataforma GitHub, (ii) percepções mensuradas por escala *Likert* e (iii) análise temática das perguntas abertas.

Indicadores de Uso

Os dados coletados das interações dos estudantes no processo de commit do GitHub são apresentados de forma consolidada na Tabela 1. O resultado das médias foi de 4,55 *commits* por estudante sem Autograder, 6,90 com Autograder por testes e 6,00 com Autograder com LLM. As duas atividades do segundo cenário (TP4 e TP5) foram agrupadas pela média individual

Tabela 1. *Commits* por estudante em cada atividade (n = 20)

Atividade	Cenário	M	DP	Mdn	IQR	Mín–Máx
TP3	Sem Autograder	4,55	4,39	4,5	1,0 - 6,0	0 - 14
TP4	Autograder por testes	6,50	6,06	5,0	2,0 - 11,0	0 - 20
TP5	Autograder por testes	7,30	8,54	5,5	2,0 - 10,0	0 - 37
TP6	Autograder com LLM	6,00	5,64	4,5	1,5 - 9,3	0 - 19

Legenda: M - média; DP - desvio-padrão; Mdn - mediana; IQR - intervalo interquartil.

A leitura das médias sugere elevação em relação ao cenário sem automação, mas três aspectos dos dados impõem cautela. Primeiro, a dispersão é elevada: em todas as atividades o desvio-padrão é próximo ou superior à média, e o intervalo observado vai de zero a 37 *commits*. Segundo, as medianas dos três cenários são idênticas (4,5 *commits*), o que indica que a elevação das médias é produzida pelas caudas da distribuição, e não por deslocamento do comportamento típico. Um único estudante concentra 25% de todos os *commits* registrados na atividade TP5; sem essa observação, a média da atividade cai de 7,30 para 5,74. Terceiro, o padrão individual é heterogêneo: na passagem do cenário sem Autograder para o cenário com testes, 12 estudantes aumentaram o número de submissões, cinco reduziram e três mantiveram; já entre o cenário com testes e o cenário com LLM, oito aumentaram e nove reduziram. Os dados são apresentados no gráfico da Figura 6.

Esses indicadores são compatíveis com maior interação com a plataforma após a introdução do Autograder, mas não sustentam, isoladamente, afirmação sobre aumento de envolvimento. Cabe registrar que o número de *commits* é sensível ao próprio desenho da ferramenta: como a correção é disparada a cada submissão, parte da variação observada pode refletir adaptação estratégica ao mecanismo (submeter para obter verificação) e não maior

dedicação à tarefa. Por essa razão, os dados de percepção apresentados a seguir constituem a evidência principal deste relato.

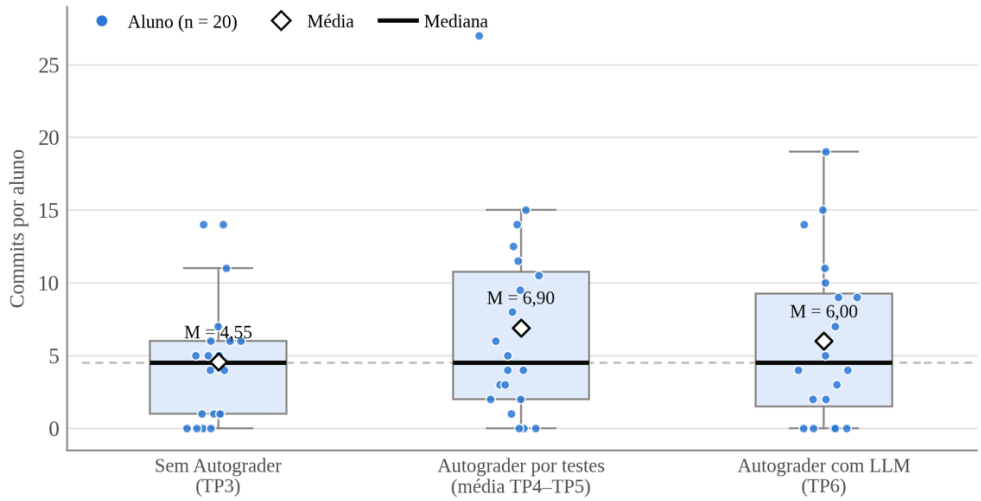


Figura 6. Distribuição de commits por estudante em cada cenário de feedback

Percepção dos estudantes

A análise dos questionários aplicados, apresentados na Tabela 2, revelou percepções majoritariamente positivas quanto à contribuição do Autograder para o processo formativo.

Tabela 2. Percepção dos estudantes por cenário de *feedback*

Construto	Sem corretor (n = 11) M (Mdn)	Autograder por testes (n = 10) M (Mdn)	Autograder com LLM (n = 9) M (Mdn)	Concordância (%)
Suficiência de orientação e feedback	2,64 (2,0)	3,10 (3,0)	3,56 (4,0)	45 → 90 → 100
Clareza do processo	2,82 (3,0)	3,40 (3,5)	3,56 (4,0)	64 → 90 → 100
Envolvimento na atividade	3,09 (3,0)	3,20 (3,5)	3,56 (4,0)	73 → 70 → 89
Agilidade na entrega	2,45 (3,0)	3,10 (3,0)	3,00 (3,0)	55 → 80 → 56
Qualidade das entregas	2,64 (3,0)	3,00 (3,0)	3,22 (4,0)	55 → 70 → 78

Escala de quatro pontos (1 = discordo fortemente; 4 = concordo fortemente). Percentuais de concordância calculados sobre respostas válidas, nas categorias "concordo em parte" e "concordo fortemente".

O cenário sem Autograder é o pior avaliado em quatro das cinco dimensões, e o padrão de melhoria acompanha a progressão dos cenários em quatro deles. O contraste mais acentuado ocorre na suficiência de orientação e feedback: a concordância passa de 45% no cenário sem

automação para 100% no cenário com LLM, com tamanho de efeito grande na comparação entre esses dois extremos. Essa é a dimensão mais diretamente ligada à função da ferramenta, e o resultado é coerente com a natureza do artefato, o Autograder atua exatamente sobre a lacuna de orientação que o cenário sem correção automatizada deixa aberta.

A agilidade na entrega é a única dimensão que não acompanha a progressão: a concordância sobe de 55% para 80% com o Autograder por testes, mas retorna a 56% com o LLM. Uma leitura possível é a de que o *feedback* gerado por modelo de linguagem, por ser mais extenso e interpretativo, demande mais tempo de leitura do que o relatório objetivo produzido pelos testes, hipótese que o desenho deste estudo não permite verificar e que fica registrada como trabalhos futuros.

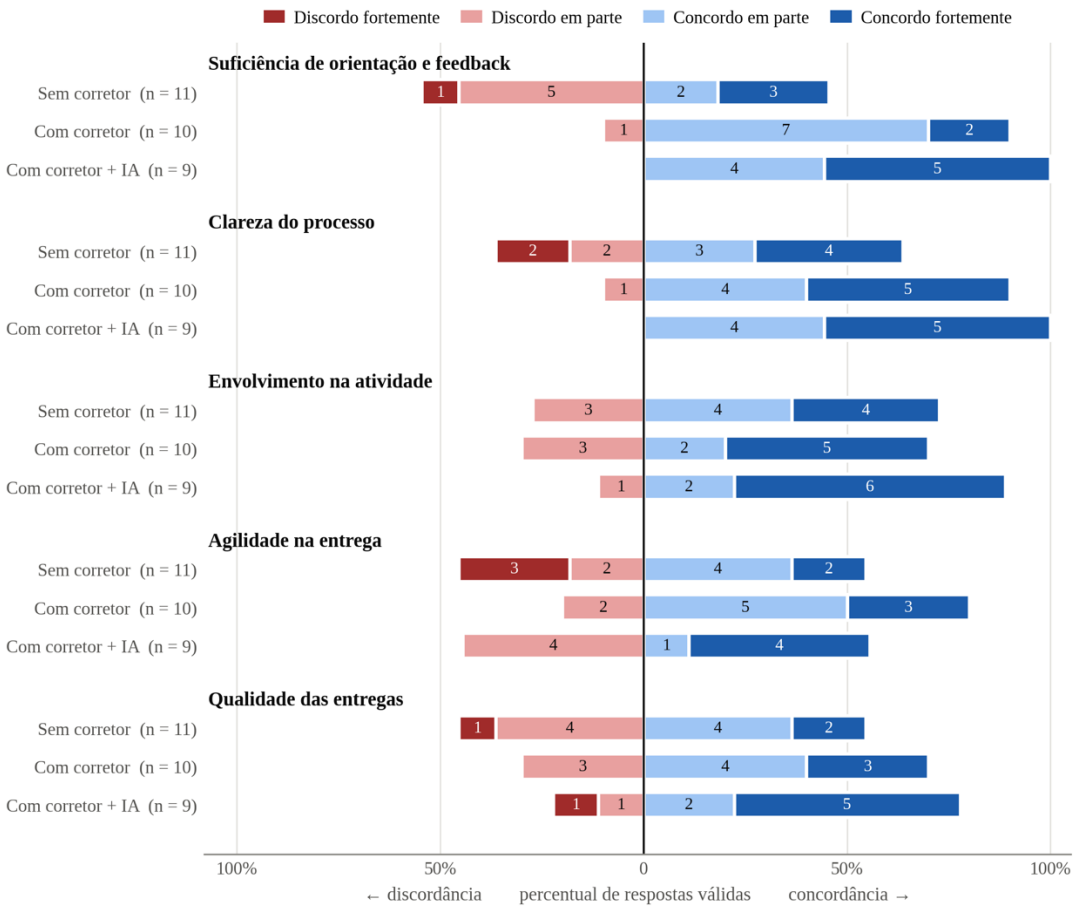


Figura 7. Distribuição das respostas por dimensão

Os resultados mostrados na Figura 7 reforçam a hipótese de que o *feedback* imediato e personalizado contribui para o aumento do envolvimento e da autonomia dos estudantes. O impacto é coerente com achados de Chan et al. (2024), que observaram ganhos significativos na qualidade das revisões de textos e no envolvimento de alunos que receberam *feedback* gerado por IA generativa. De forma semelhante, em um estudo exploratório com um grupo reduzido de participantes, Kusam (2024) reporta uma tendência de que a integração de *Large Language Models (LLMs)* em ambientes de *Project-Based Learning* potencializa a autorregulação dos estudantes e reduz o tempo de resposta do professor.

Os dados da avaliação geral do Autograder são apresentados na Tabela 3 e Figura 8. Nenhum respondente assinalou “discordo fortemente” em qualquer dos quatro itens. A identificação de erros não percebidos pelo próprio estudante reúne a maior concentração de concordância forte, oito dos dez respondentes válidos, o que situa a contribuição mais reconhecida da ferramenta no nível da tarefa descrito por Hattie e Timperley (2007): informar o estudante sobre a correção do que foi produzido e sobre o que precisa ser ajustado. Os níveis de processo e de autorregulação, que os autores associam a efeitos mais duradouros, dependem de mediação docente que a ferramenta não substitui.

Tabela 3. Avaliação geral do Autograder

Afirmação	n válido	DF	DP	CP	CF	Concordância
Incentivou a revisar e refazer submissões	10	0	0	4	6	100%
Feedback claro e útil para a aprendizagem	10	0	1	6	3	90%
Ajudou a identificar erro não percebido	10	0	0	2	8	100%
Fez sentir mais confiante para resolver sozinho	9	0	1	5	3	89%

Legenda: DF = discordo fortemente; DP = discordo em parte; CP = concordo em parte; CF = concordo fortemente. Respostas "não consigo responder" (uma a duas por item) excluídas do cálculo.

Registre-se ainda um dado que qualifica os demais: embora nove dos onze respondentes declarassem interesse alto pela temática no início da disciplina, apenas seis mantinham esse nível ao final, com quatro em nível médio e um em nível baixo. A avaliação positiva da ferramenta, portanto, convive com uma trajetória de interesse que não é uniformemente ascendente.

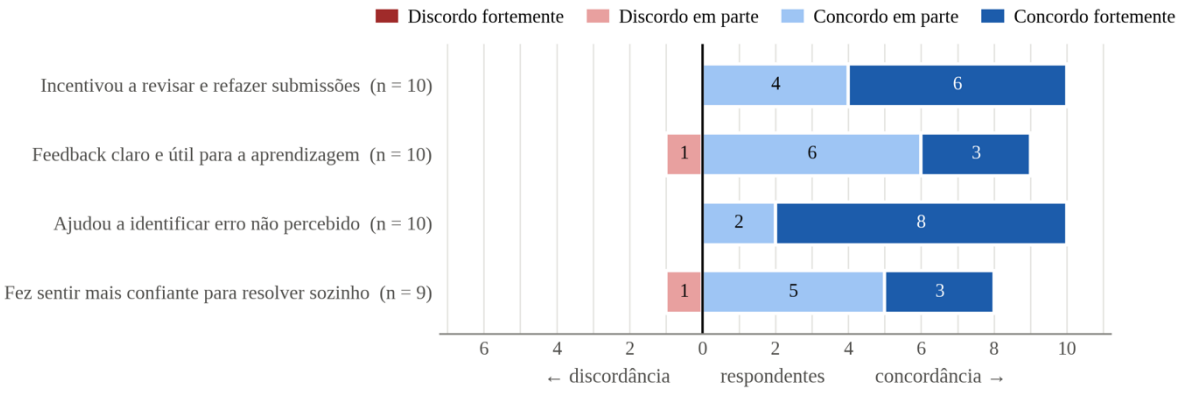


Figura 8. Avaliação geral do Autograder pelos estudantes

Análise das perguntas abertas

A análise qualitativa das respostas abertas permitiu compreender aspectos subjetivos associados ao uso do Autograder. A questão aberta foi respondida por sete dos onze participantes. Os relatos dos participantes indicaram três eixos interpretativos principais: (i) clareza e detalhamento das orientações, (ii) sensação de autonomia e autoconfiança, e (iii) motivação e envolvimento com a tarefa.

No primeiro eixo, os alunos destacaram que o *feedback* automático foi mais direto e objetivo, fornecendo informações específicas sobre erros e sugerindo formas de correção. Um dos respondentes afirmou: "A correção automática agiliza a percepção do erro e induz a busca por mais conhecimento em determinado assunto."

No segundo eixo, a percepção de autonomia e autoconfiança emergiu como um dos resultados mais relevantes. Diversos estudantes relataram que a possibilidade de testar e submeter novamente o código sem depender do retorno do professor gerou maior senso de controle sobre a própria aprendizagem.

Por fim, no eixo da motivação e envolvimento, as respostas evidenciaram que o caráter imediato do *feedback* atuou como reforço positivo, estimulando o esforço contínuo. Um aluno sintetizou essa percepção ao declarar: "Achei muito interessante a correção automática, além da resposta ser bem rápida, ela ajuda bastante a ver erros ou coisas que estão faltando na nossa atividade."

Embora a maioria das percepções tenha sido positiva, alguns alunos apontaram limitações na interpretação do *feedback* em casos mais complexos, especialmente na fase de uso do LLM, em que o retorno era mais detalhado, porém, ocasionalmente, impreciso. Essa observação reforça a necessidade de acompanhamento docente e curadoria pedagógica no uso de modelos generativos, conforme Kasneci et al. (2023) ao enfatizarem os riscos de dependência e enviesamento em *feedbacks* automatizados.

Síntese interpretativa dos resultados

De modo geral, os resultados obtidos confirmam parcialmente as hipóteses formuladas no início da pesquisa. Verificou-se que o *feedback* imediato e estruturado trouxe uma percepção de suficiência e clareza das informações para uma experiência formativa mais dinâmica. Além disso, o desenho da solução transfere para o Autograder as tarefas de correção que, até então, são executadas manualmente, automatizando um trabalho do docente e oferecendo retorno individualizado ao aluno em tempo real. No entanto, o experimento não inclui o acompanhamento de indicadores do esforço dos docentes envolvidos, não permitindo afirmar sobre a redução efetiva da sobrecarga a partir dos dados apresentados.

Os dados também indicam que o sistema pode atender a diferentes perfis de estudantes, independentemente do nível de conhecimento prévio, embora alunos com maior familiaridade técnica tenham relatado menor dependência do *feedback* automatizado. Raihan et al. (2025) registram que a literatura sobre LLMs na educação de computação se distribui de forma desigual entre níveis de ensino, o que reforça a necessidade de investigar esse aspecto com maior profundidade.

Portanto, os achados deste estudo evidenciam que o uso de mecanismos automatizados de *feedback* constitui uma estratégia viável para promover escalabilidade, personalização e envolvimento em jornadas formativas, desde que acompanhada por diretrizes pedagógicas claras e por um design instrucional que valorize o papel ativo do estudante.

6. Conclusões e Recomendações

O presente estudo teve como propósito analisar a implementação e os efeitos de um sistema de *feedback* automatizado aplicado a atividades práticas de uma disciplina de Engenharia de Software, no contexto de jornadas formativas voltadas ao desenvolvimento de competências técnicas e comportamentais. A experiência relatada evidencia que a introdução de mecanismos de correção automática, integrados à infraestrutura do GitHub Classroom e apoiados em modelos de linguagem de larga escala (LLMs), pode gerar ganhos em agilidade, envolvimento e personalização do processo de aprendizagem.

Os resultados indicam que o *feedback* imediato, especialmente aquele que combina informações objetivas dos testes de software com interpretações geradas por modelos de linguagem, exerce impacto positivo na percepção dos estudantes quanto à clareza na compreensão dos erros e a autoconfiança dos estudantes. Essa constatação reforça o que Hattie e Timperley (2007) identificam como o "poder do *feedback*", capaz de reduzir o intervalo entre o desempenho atual e o estabelecido como meta por meio de informações precisas e imediatas sobre a aprendizagem.

Além disso, a análise das percepções qualitativas revelou que os alunos se sentiram mais autônomos e motivados ao poder iterar rapidamente sobre seus próprios códigos, utilizando o *feedback* como instrumento de autoavaliação. Tal comportamento está em consonância com as diretrizes de *Competency-Based Education (CBE)* discutidas por Levine e Patrick (2019), que enfatizam a importância de avaliações contínuas, baseadas em evidências e orientadas ao domínio de competências.

Do ponto de vista pedagógico, o Autograder mostrou-se uma ferramenta de apoio à docência, com potencial para aliviar a carga operacional de correção e ampliar o alcance do acompanhamento em turmas com um volume maior de estudantes. Esse potencial está associado à proposta da solução que permite o disparo da correção a partir da sinalização de entrega (via commit no repositório do projeto) sem intervenção humana e apresentação de um relatório consolidado. Embora ainda não tenha sido possível avaliar a qualidade da orientação em escala, essa característica posiciona o Autograder como um recurso estratégico para instituições que enfrentam o desafio de oferecer formação personalizada em larga escala, especialmente em programas de capacitação tecnológica e educação continuada.

No entanto, os achados também indicam limitações e precauções importantes. Trata-se de relato de experiência conduzido em uma única turma, com 20 participantes, sem grupo de controle e sem randomização, o que impede atribuição causal entre a introdução do Autograder e as variações observadas. Os cenários também se sucedem no tempo, de modo que efeitos de maturação e de familiaridade crescente com a disciplina não podem ser separados do efeito da ferramenta. Adicionalmente, o uso de modelos generativos requer supervisão docente e curadoria pedagógica, uma vez que o *feedback* fornecido por LLMs pode conter imprecisões ou interpretações enviesadas, o que é amplamente discutido por Kasneci et al. (2023) ao tratarem da responsabilidade ética e cognitiva no uso de IA em contextos educacionais. Assim, a integração entre algoritmos e ação docente deve ser compreendida como complementar e não substitutiva, preservando o papel do professor como mediador da aprendizagem. Por fim, não foram coletados dados sobre o trabalho docente, o desempenho acadêmico ou a retenção de aprendizagem. Afirmações sobre redução de carga de trabalho do professor, sobre viabilidade de turmas maiores e sobre eficácia formativa dependem dessas dimensões e permanecem, neste trabalho, como propostas de pesquisas futuras.

A partir das evidências observadas, recomenda-se:

1. Planejamento pedagógico orientado ao *feedback* - o uso de ferramentas automáticas pode ser incorporado desde o design instrucional da disciplina, com objetivos claros e critérios de avaliação compatíveis com os resultados esperados;

2. Integração gradual de tecnologias de IA - recomenda-se iniciar com *feedbacks* baseados em testes automatizados (objetivos) antes de incorporar modelos de linguagem (subjetivos), garantindo consistência e previsibilidade nas respostas;

3. Acompanhamento docente sistemático - os professores devem monitorar periodicamente os relatórios gerados pelo sistema, intervindo quando o *feedback* automatizado se mostrar ambíguo ou excessivamente genérico;

4. Promoção da metacognição e da autoavaliação - atividades formativas devem estimular os alunos a interpretar criticamente o *feedback* recebido, identificando padrões de erro e estratégias de melhoria contínua; e

5. Formação de docentes e tutores - é essencial capacitar professores no uso ético, técnico e pedagógico de ferramentas de IA, assegurando o alinhamento entre as práticas de *feedback* automatizado e os princípios de avaliação formativa.

Para a consolidação de evidências sobre o potencial dos sistemas de *feedback* automatizado, recomenda-se que estudos futuros:

- ampliem o número e a diversidade de participantes, contemplando diferentes cursos e níveis de experiência com programação;

- avaliem longitudinalmente o impacto do *feedback* automatizado sobre indicadores de aprendizagem e desempenho acadêmico;

- investiguem comparativamente a efetividade de diferentes modelos de linguagem e estratégias de personalização do *feedback*;

- explorem a integração de ferramentas de *feedback* automático com painéis de acompanhamento docente e mecanismos adaptativos de recomendação, alinhados ao perfil de cada aluno;

- analisem dimensões éticas e psicológicas do uso de IA em processos avaliativos, com foco na confiança, transparência e aceitabilidade das respostas geradas.

Declaração Sobre o Uso de Inteligência Artificial.

Os autores informam que recorreram a ferramentas de inteligência artificial baseadas em modelos de linguagem de larga escala (LLMs) exclusivamente para apoio na revisão de estilo, na sugestão de organização textual e apoio na geração de gráficos. Todas as análises, interpretações e conclusões apresentadas são de responsabilidade integral dos autores.

7 Referências Bibliográficas

Chan, S., Lo, N., & Wong, A. (2024). Generative AI and essay writing: Impacts of automated feedback on revision performance and engagement. *rEFLections, 31*(3), 1249–1284.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.

Gervais, J. (2016). The operational definition of competency-based education. *The Journal of Competency-Based Education, 1*(2), 98–106. https://doi.org/10.1002/cbe2.1011

Guerra Hahn, M., Baldiris Navarro, S. M., de la Fuente Valentín, L., & Burgos, D. (2021). A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access, 9*, 108190–108198. https://doi.org/10.1109/ACCESS.2021.3100890

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kusam, V. A. (2024). *Generative-AI assisted feedback provisioning for project-based learning in CS education* [Dissertação de mestrado não publicada]. University of Michigan–Dearborn.

Levine, E., & Patrick, S. (2019). *What is competency-based education? An updated definition*. Aurora Institute.

Mukul, E., & Büyüközkan, G. (2023). Digital transformation in education: A systematic review of education 4.0. *Technological Forecasting and Social Change, 194*, 122664. https://doi.org/10.1016/j.techfore.2023.122664

Oroszi, T. (2020). Competency-based education. *Creative Education, 11*(11), 2467–2476. https://doi.org/10.4236/ce.2020.1111181

Raihan, N., Siddiq, M. L., Santos, J. C. S., & Zampieri, M. (2025). Large language models in computer science education: A systematic literature review. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1* (pp. 938–944). Association for Computing Machinery. https://doi.org/10.1145/3641554.3701863