

Aprender ou Delegar? Um Estudo sobre Pensamento Crítico no Uso de IA Generativa em Avaliação Heurística

Fabiann Matthaues^{1,2}, Luana Rodrigues², Matheus Lima², Tayana Conte¹

¹Pós-Graduação em Informática (PPGI) – Universidade Federal do Amazonas (UFAM)

Av. General Rodrigo Otávio, Coroadó, Manaus – AM – Brasil

²Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM)
Campus Manaus Zona Leste – Manaus – AM – Brasil

{fabiann.dantas, 2023005010, 2023010727}@ifam.edu.br
tayana@icomp.ufam.edu.br

Abstract. *This study investigates the use of Generative Artificial Intelligence (GenAI) by students in Heuristic Evaluation activities, analyzing its impacts on usability problem identification and the critical use of generated responses. An experiment was conducted with 24 Software Engineering students divided into a control group (without AI) and an experimental group (with AI). Quantitative results showed no statistically significant difference between the groups, although the AI group achieved a lower average accuracy rate. Qualitative analysis and interaction logs revealed a discrepancy between students' perceptions and their actual use of AI, highlighting predominantly passive usage, uncritical reproduction of responses, hallucinations, sycophancy, and difficulty identifying AI-generated errors. The findings suggest that the use of GenAI without proper pedagogical mediation may encourage cognitive delegation processes in conceptual learning activities.*

Resumo. *Este trabalho investiga o uso de Inteligência Artificial Generativa (IA-Gen) por estudantes em atividades de Avaliação Heurística, analisando tanto a identificação de problemas de usabilidade quanto o pensamento crítico no uso das respostas geradas pela IA. Foi conduzido um experimento com 24 alunos de Engenharia de Software, divididos em grupo controle (sem IA) e experimental (com IA). Os resultados quantitativos não indicaram diferença estatisticamente significativa entre os grupos, embora o grupo com IA tenha apresentado menor média de acertos. A análise qualitativa e dos logs revelou discrepância entre a percepção dos estudantes e o uso real da IA, evidenciando predominância de uso passivo, reprodução acrítica de respostas, alucinações, sicofantia e dificuldade em identificar erros gerados pela ferramenta. Os resultados sugerem que o uso de IAGen sem mediação pedagógica adequada pode favorecer processos de delegação cognitiva em atividades conceituais de aprendizagem.*

1. Introdução

O avanço da Inteligência Artificial Generativa (IAGen), tem transformado diferentes contextos educacionais [Wang et al. 2025]. Ferramentas baseadas nessa tecnologia vêm sendo amplamente utilizadas por estudantes para apoiar a aprendizagem, a

resolução de problemas e atividades acadêmicas [Yan et al. 2024]. Na Educação em Computação, sua adoção tem crescido devido à capacidade de fornecer respostas imediatas, explicações contextualizadas e suporte interativo durante o processo de aprendizagem [Becker et al. 2023, Branco et al. 2025].

Apesar de seu potencial educacional, a IAGen também aponta preocupações relacionadas à dependência excessiva dos estudantes [De Cremer et al. 2023, Mittal et al. 2024]. A literatura destaca riscos como redução do esforço cognitivo [Gerlich 2025], aceitação acrítica de respostas [Favero et al. 2025], dificuldades de validação das informações geradas [Bauschard and Quidwai 2024] e delegação cognitiva [Hong et al. 2025]. Esses desafios tornam-se relevantes em atividades que exigem análise crítica e tomada de decisão [Chan and Hu 2023].

No contexto da Interação Humano-Computador (IHC), a Avaliação Heurística requer que estudantes identifiquem problemas de usabilidade, interpretem heurísticas e justifiquem suas decisões [Sobrino-Duque et al. 2022, Barbosa et al. 2021]. Trata-se de uma atividade essencialmente analítica, na qual o estudante precisa relacionar evidências observadas na interface a conceitos teóricos de usabilidade [Sobrino-Duque et al. 2022].

Nesse cenário, a utilização de IAGen apresenta tanto oportunidades quanto desafios. Por um lado, a ferramenta pode apoiar o raciocínio e auxiliar na compreensão das heurísticas [Dibia et al. 2023], por outro, pode favorecer a aceitação passiva de respostas e a dependência tecnológica durante a atividade [Barke et al. 2023]. Embora estudos recentes tenham investigado o uso de IAGen em diferentes atividades educacionais, ainda são limitadas as evidências sobre como estudantes utilizam essas ferramentas em tarefas que exigem análise crítica e tomada de decisão [Ravšelj et al. 2025].

Diante desse contexto, este trabalho investiga o uso de IAGen por estudantes em atividades de Avaliação Heurística, analisando tanto o desempenho na identificação de problemas de usabilidade quanto o pensamento crítico no uso das respostas produzidas pela IAGen. Para isso, foi conduzido um experimento controlado com estudantes de Engenharia de Software, divididos entre grupo controle (sem IAGen) e grupo experimental (com IAGen). O estudo combinou a análise dos resultados produzidos pelos participantes, suas percepções sobre a atividade e os registros de interação com a IAGen, buscando compreender padrões de uso, estratégias de validação das respostas e comportamentos que possam influenciar a autonomia dos estudantes durante o processo de aprendizagem.

Como contribuições, este trabalho apresenta: (i) evidências empíricas sobre o uso de IAGen em atividades de Avaliação Heurística; (ii) uma análise dos padrões de interação com a IA e suas relações com o pensamento crítico durante atividades conceituais de aprendizagem; e (iii) discussões sobre implicações pedagógicas relacionadas à mediação do uso de IAGen em atividades conceituais de aprendizagem.

2. Referencial Teórico

2.1. Inteligência Artificial Generativa na Educação

A Inteligência Artificial Generativa (IAGen), especialmente por meio de modelos de linguagem de grande escala (LLMs), tem transformado os processos de ensino, aprendizagem e avaliação na educação [Silvestre et al. 2023]. Essas tecnologias vêm sendo utilizadas como suporte à aprendizagem, oferecendo explicações adaptadas,

auxílio à programação, produção textual e apoio à realização de atividades conceituais [Bahroun et al. 2023].

Entre seus principais potenciais, destacam-se a personalização da aprendizagem, a oferta de *feedback* automatizado e o apoio à aprendizagem autorregulada [Xia et al. 2024]. Além disso, ferramentas de IAGen podem atuar como mediadoras pedagógicas, auxiliando estudantes na resolução de dúvidas e no desenvolvimento de atividades educacionais [Triberti et al. 2024]. Apesar dessas oportunidades, a adoção da IAGen também traz desafios éticos, pedagógicos e cognitivos [Baidoo-Anu and Ansah 2023]. Nesse contexto, destaca-se a necessidade de repensar métodos de avaliação, promover a alfabetização em IA e capacitar professores para o uso crítico dessas tecnologias [Sousa and Cruz 2024].

2.2. Avaliação Heurística

A usabilidade é um atributo de qualidade relacionado à facilidade com que usuários interagem com um sistema, influenciando diretamente a experiência de uso [Nielsen 1994a, Natella et al. 2018]. Nesse contexto, a inspeção de usabilidade constitui um método amplamente utilizado para identificação de problemas em interfaces, destacando-se a Avaliação Heurística como uma das técnicas mais adotadas [Nielsen 1994b, Rocha and Baranauskas 2003].

A Avaliação Heurística consiste na inspeção sistemática de uma interface com base em princípios de usabilidade previamente estabelecidos, denominados heurísticas [Nielsen 1994b]. Entre as abordagens mais difundidas está a proposta de Nielsen, composta por dez princípios gerais de usabilidade¹. Durante a avaliação, os participantes analisam a interface em busca de problemas de usabilidade e os associam às heurísticas correspondentes, produzindo justificativas para suas decisões [Nielsen 1994b]. Por sua simplicidade e baixo custo, a Avaliação Heurística é amplamente utilizada no ensino de IHC, exigindo interpretação conceitual e tomada de decisão [Rocha and Baranauskas 2003].

3. Trabalhos Relacionados

O estudo de [Guerino et al. 2025] comparou o desempenho do GPT-4 com o de especialistas em IHC na realização de avaliações heurísticas. Os resultados mostraram que o modelo identificou apenas 21,2% dos problemas encontrados pelos avaliadores humanos, além de produzir falsos positivos associados a alucinações. Com base nisso, os autores recomendam o uso de LLMs como ferramentas de apoio ao processo analítico, e não como substitutos do julgamento humano.

No trabalho de [Zhong et al. 2025], modelos multimodais foram empregados em avaliações heurísticas sintéticas, mas apresentaram limitações em relação às avaliações realizadas por especialistas humanos. Apesar dessas contribuições, tanto esse estudo quanto o de [Guerino et al. 2025] concentram-se no desempenho dos modelos, sem investigar como os usuários interagem com as respostas produzidas pela IA durante a atividade.

Na educação superior, [Sajja et al. 2025] analisou como estudantes interagem com ferramentas de IA focando em percepções de usabilidade, confiança, ética e valor educacional. Os alunos consideram assistentes de IA acessíveis e confortáveis, principalmente

¹<https://nngroup.com/articles/ten-usability-heuristics>

para tarefas e compreensão de conceitos, mas os destacam como complemento, não substituto da instrução humana, levantando dúvidas éticas e de política institucional.

Portanto, o presente trabalho diferencia-se dos estudos analisados ao investigar o uso da IAGen por estudantes em atividades de Avaliação Heurística, analisando o desempenho na identificação de problemas de usabilidade. Esta pesquisa busca compreender como os estudantes utilizam, validam e incorporam as respostas produzidas pela IA durante a realização da atividade.

4. Metodologia

Este estudo foi conduzido por meio de um experimento controlado com o objetivo de investigar o uso de IAGen por estudantes em atividades de Avaliação Heurística no contexto do ensino de IHC. Especificamente, a pesquisa analisou três aspectos: (i) o desempenho dos estudantes na identificação de problemas de usabilidade; (ii) suas percepções sobre aprendizagem e realização da atividade; e (iii) os padrões de interação com a IAGen, buscando identificar evidências de pensamento crítico durante o uso da ferramenta. Para orientar a condução do estudo, foram definidas as seguintes questões de pesquisa.

- **QP1:** O uso de IAGen influencia a capacidade dos estudantes de identificar problemas de usabilidade durante atividades de Avaliação Heurística?
- **QP2:** Qual o impacto do uso de IAGen na percepção de aprendizagem dos estudantes?
- **QP3:** Como os estudantes utilizam as respostas produzidas pela IAGen durante a atividade e quais evidências de pensamento crítico podem ser observadas em seus padrões de interação?

Para responder à QP1, foram formuladas as seguintes hipóteses estatísticas:

- **H0:** Não existe diferença significativa no desempenho entre estudantes que utilizam IAGen e estudantes que realizam a atividade sem o auxílio da ferramenta.
- **H1:** Existe diferença significativa no desempenho entre estudantes que utilizam IAGen e estudantes que realizam a atividade sem o auxílio da ferramenta.

As questões QP2 e QP3 possuem caráter exploratório e foram investigadas por meio dos questionários de percepção e dos registros de interação com a IAGen coletados durante o experimento.

4.1. Contexto

O experimento foi conduzido em uma disciplina de IHC de um curso de graduação em Engenharia de Software. Segundo [Wohlin et al. 2012], estudos experimentais são importantes para compreender os impactos de novas tecnologias em atividades específicas. Nesse contexto, este trabalho investiga o uso da IAGen como ferramenta de apoio em atividades de Avaliação Heurística, uma tarefa que exige identificação de problemas de usabilidade, interpretação conceitual e tomada de decisão [Sobrinho-Duque et al. 2022].

Por sua natureza educacional, o estudo foi realizado em ambiente de sala de aula, permitindo observar a interação dos estudantes com a IAGen em um cenário autêntico de aprendizagem [Zhan et al. 2021]. Além do desempenho na atividade, buscou-se compreender como os estudantes utilizam as respostas produzidas pela IA durante seu processo de análise e resolução da tarefa.

4.2. Seleção de Variáveis

A variável independente do estudo foi o uso de IAGen durante a realização da atividade de Avaliação Heurística, definida em dois tratamentos: sem IA (conforme adotado no Grupo controle (GC)) e com IA (conforme adotado no Grupo Experimental (GE)). As variáveis dependentes foram definidas de acordo com os objetivos da pesquisa. O desempenho dos estudantes foi avaliado por meio do número de problemas de usabilidade identificados durante a atividade, da quantidade de acertos na associação entre problemas e heurísticas e do número de erros conceituais observados nas respostas produzidas.

Além das métricas de desempenho, foram analisadas as percepções dos participantes sobre a atividade e sobre o uso da IAGen. Para os participantes do GE, também foram investigados os padrões de interação com a ferramenta por meio dos registros das conversas realizadas durante a atividade. Esses registros foram utilizados para identificar estratégias de uso, evidências de validação das respostas produzidas pela IA e aspectos relacionados ao pensamento crítico.

4.3. Seleção dos Participantes

O estudo contou com a participação voluntária de 24 estudantes de graduação em Engenharia de Software matriculados na disciplina de IHC. Todos os participantes assinaram o Termo de Consentimento Livre e Esclarecido (TCLE) e o estudo foi aprovado por um Comitê de Ética em Pesquisa sob CAAE nº 94868425.0.0000.5020 e Parecer nº 8.167.870/2026.

Antes da execução do experimento, os estudantes responderam a um questionário de caracterização para identificar sua experiência prévia com usabilidade, Avaliação Heurística e ferramentas de IAGen. Com base nessas informações, os participantes foram distribuídos entre os grupos por meio de amostragem aleatória estratificada [Athey and Imbens 2017], buscando reduzir possíveis efeitos da experiência prévia sobre os resultados. A participação de estudantes permitiu conduzir a avaliação no público-alvo da atividade investigada.

4.4. Design Experimental

O design adotado neste estudo foi de um fator com dois tratamentos. O fator corresponde ao uso de Inteligência Artificial Generativa (IAGen) durante a realização da atividade de Avaliação Heurística. Os tratamentos foram definidos como: (i) realização da atividade sem o auxílio de IAGen, correspondente ao Grupo Controle (GC), e (ii) realização da atividade com acesso livre à IAGen, correspondente ao Grupo Experimental (GE).

A distribuição dos participantes entre os grupos foi realizada de forma aleatória após a aplicação do questionário de caracterização. Como os participantes apresentaram perfis semelhantes em relação à experiência prévia com Avaliação Heurística e ferramentas de IA, não foi necessária a aplicação de técnicas adicionais de balanceamento entre os tratamentos. Cada grupo foi composto por 12 participantes e cada estudante foi submetido a apenas um tratamento. Dessa forma, o estudo adotou um desenho experimental entre sujeitos *Between-subjects design* [Lazar et al. 2017], evitando possíveis efeitos de aprendizagem decorrentes da exposição dos participantes a múltiplas condições experimentais. A Figura 1 apresenta uma visão geral do desenho experimental adotado neste estudo.

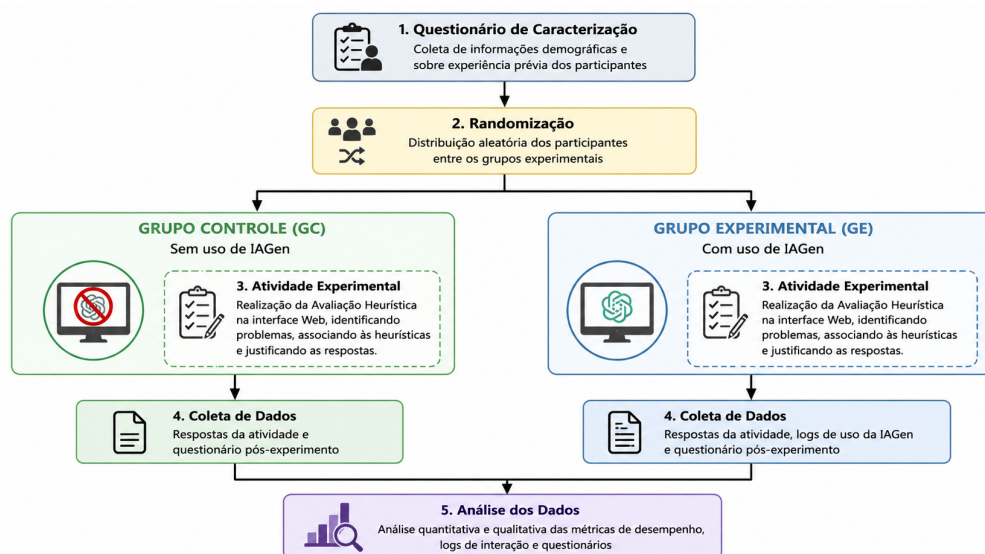


Figura 1. Visão geral do desenho experimental adotado no estudo.

4.5. Instrumentação

Para apoiar a condução do experimento, foram preparados diferentes artefatos, todos disponibilizados no material suplementar do estudo. Inicialmente, foi utilizado um Termo de Consentimento Livre e Esclarecido (TCLE). Também foi aplicado um formulário de caracterização destinado ao levantamento do perfil dos participantes.

A atividade experimental foi conduzida utilizando uma interface Web previamente selecionada, que serviu como base para a construção do *gold standard* utilizado na análise dos resultados. Os participantes registraram os problemas identificados, as heurísticas associadas e suas justificativas por meio de um *template* estruturado. Para o Grupo Experimental, foi disponibilizado o ChatGPT (GPT-5.5), escolhido por representar uma das ferramentas de IAGen mais amplamente utilizadas por estudantes em contextos educacionais. Ao término da atividade, os participantes responderam a um questionário pós-experimento e foram coletados os logs de interação dos participantes do GE.

4.6. Preparação e Execução

Previamente à aplicação do estudo, foi construído um *gold standard* por meio de uma validação sistemática conduzida por três especialistas em IHC, que realizaram de forma independente a Avaliação Heurística da interface e posteriormente discutiram divergências até alcançar consenso sobre os problemas de usabilidade, heurísticas associadas e respectivas justificativas. Também foram realizados testes piloto com dois estudantes concluintes de Engenharia de Software para avaliar a clareza das instruções, a adequação dos instrumentos e o tempo necessário para a atividade, resultando em ajustes nos materiais utilizados.

Antes do experimento, todos os participantes receberam uma aula teórica de aproximadamente 60 minutos sobre as heurísticas de Nielsen e os fundamentos da Avaliação Heurística. Em seguida, foi realizada uma atividade prática com correção e discussão coletiva das respostas para consolidar os conceitos e familiarizar os estudantes com o processo de inspeção de interfaces. Em um encontro posterior da disciplina, os parti-

participantes foram distribuídos entre Grupo Controle (GC) e Grupo Experimental (GE). A atividade foi realizada individualmente, com duração máxima de 60 minutos, durante o horário regular da disciplina. Após receberem as instruções gerais e acesso à interface Web selecionada, os estudantes realizaram uma Avaliação Heurística registrando os problemas identificados, as heurísticas associadas e suas justificativas em um *template* disponibilizado pelos pesquisadores.

Os participantes do GC realizaram a tarefa sem suporte de IA, utilizando apenas os conhecimentos adquiridos durante as aulas. Já os participantes do GE tiveram acesso livre ao ChatGPT (GPT-5.5), podendo utilizá-lo durante toda a atividade para apoiar suas análises e tomadas de decisão. Durante a execução, os pesquisadores permaneceram disponíveis apenas para esclarecimentos procedimentais, sem fornecer orientações sobre os problemas presentes na interface ou sobre as respostas esperadas.

Ao término, todos os participantes responderam a um questionário pós-experimento contendo questões em escala Likert e questões abertas relacionadas à percepção de aprendizagem, dificuldades encontradas e experiência de uso. Adicionalmente, os participantes do GE responderam questões específicas sobre a utilização da IAGen. Os registros das interações realizadas com o ChatGPT também foram coletados para posterior análise.

5. Análise dos Resultados

5.1. Análise do Desempenho dos Participantes

O desempenho dos participantes foi analisado sob duas perspectivas complementares: (i) o número total de problemas de usabilidade identificados e (ii) o número de acertos em relação ao *gold standard* elaborado pelos especialistas. Para cada participante, os problemas reportados foram comparados ao conjunto de problemas previamente validados, permitindo distinguir problemas corretamente identificados de erros conceituais e classificações inadequadas.

A análise estatística foi conduzida utilizando *Python 3.12* e as bibliotecas *Pandas* e *SciPy*. Inicialmente, foi aplicado o teste de normalidade de Shapiro-Wilk. Como os dados não apresentaram distribuição normal, optou-se pela utilização do teste não paramétrico de Mann-Whitney para comparação entre os grupos.

Os participantes do Grupo Controle (GC) apresentaram mediana de 7 problemas identificados ($Q_1 = 6,0$; $Q_3 = 7,5$), enquanto os participantes do Grupo Experimental (GE) obtiveram mediana de 6 problemas ($Q_1 = 4,8$; $Q_3 = 6,3$). A comparação entre os grupos não revelou diferença estatisticamente significativa para o número de problemas identificados ($U = 95,0$, $p = 0,183$), não sendo possível rejeitar a hipótese nula associada à QP1. O tamanho de efeito observado foi pequeno a moderado ($r = 0,27$), indicando que a magnitude da diferença entre os grupos foi limitada.

Quando analisado o número de acertos em relação ao *gold standard*, os participantes do GC apresentaram mediana de 5 acertos ($Q_1 = 4,0$; $Q_3 = 6,0$), enquanto os participantes do GE obtiveram mediana de 3,5 acertos ($Q_1 = 3,0$; $Q_3 = 4,0$). A comparação entre os grupos também não revelou diferença estatisticamente significativa ($U = 101,0$, $p = 0,095$), embora tenha sido observada uma tendência de melhor desempenho do GC. O tamanho de efeito observado foi pequeno a moderado ($r = 0,34$).

A Figura 2 apresenta a distribuição dos resultados para ambas as métricas. Observa-se que o GC apresentou valores centrais superiores tanto para o número total de problemas identificados quanto para o número de acertos. Embora essas diferenças não tenham atingido significância estatística, os resultados indicam uma tendência consistente de desempenho superior dos estudantes que realizaram a atividade sem o auxílio da IAGen.

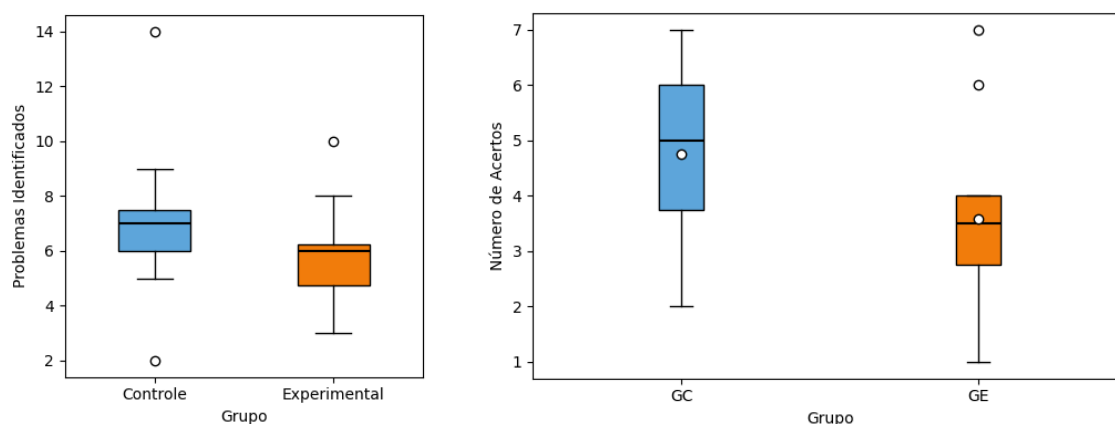


Figura 2. Distribuição dos resultados de desempenho dos participantes.

Complementarmente, a Tabela 1 apresenta as médias das principais métricas analisadas durante a atividade. Os participantes do GC identificaram, em média, mais problemas de usabilidade e obtiveram maior número de acertos quando comparados aos participantes do GE. Por outro lado, a média de erros conceituais foi equivalente entre os grupos. Em conjunto, esses resultados reforçam a tendência observada nas distribuições apresentadas na Figura 2, indicando desempenho superior do GC, embora sem evidências estatísticas suficientes para caracterizar uma diferença significativa entre os tratamentos.

Tabela 1. Médias das métricas de desempenho

Métrica	GC	GE
Problemas identificados	7,08	5,92
Acertos	4,75	3,58
Erros conceituais	2,33	2,33

5.2. Percepções dos Estudantes

Após o experimento, os participantes responderam a um questionário contendo questões em escala Likert e questões abertas relacionadas à aprendizagem, compreensão das heurísticas, esforço mental, confiança nas respostas e uso da IAGen. O instrumento completo encontra-se disponibilizado nos artefatos suplementares do estudo.

5.2.1. Percepções sobre a atividade

Para analisar as percepções dos participantes sobre a atividade, foi calculada a porcentagem de respostas positivas para cada afirmação do questionário, considerando as opções 4 ("Concordo Parcialmente") e 5 ("Concordo Totalmente") da escala Likert. A Tabela 2

apresenta os resultados obtidos para os grupos Controle e Experimental, permitindo comparar aspectos relacionados à aprendizagem, compreensão das heurísticas, facilidade de execução da atividade, esforço mental e confiança nas respostas fornecidas.

Tabela 2. Distribuição percentual de respostas positivas (4 e 5)

Questão	GC	GE
Q1 - Aprendizagem Percebida	100%	92,3%
Q2 - Compreensão das heurísticas	84,6%	84,6%
Q3 - Facilidade de identificação	61,5%	76,9%
Q4 - Associação problema-heurística	46,2%	69,2%
Q5 - A atividade exigiu muito esforço mental	38,5%	15,4%
Q6 - Dificuldade em compreender os problemas encontrados	53,8%	84,6%
Q7 - Confiança nas respostas fornecidas	61,5%	92,3%
Q8 - Certeza que as respostas estavam corretas	69,2%	69,2%

De modo geral, ambos os grupos relataram percepções positivas sobre a atividade, especialmente nos aspectos relacionados à aprendizagem percebida (Q1) e à compreensão das heurísticas (Q2). Os resultados indicam que os participantes perceberam a atividade como útil para apoiar o aprendizado dos conceitos de Avaliação Heurística, independentemente do tratamento adotado.

Ao comparar os grupos, observa-se que os participantes do GE apresentaram maior concordância nas questões relacionadas à facilidade para identificar problemas de usabilidade (Q3), à associação entre problemas e heurísticas (Q4) e à confiança nas respostas fornecidas (Q7 e Q8), sugerindo que a utilização da IAGen contribuiu para uma percepção mais positiva da atividade e para um maior senso de segurança durante sua execução.

Em relação ao esforço mental e às dificuldades encontradas, observa-se um comportamento distinto entre os grupos. Os participantes do GE relataram menor esforço mental durante a execução da atividade (Q5), o que pode estar relacionado ao suporte fornecido pela IAGen na interpretação das interfaces e na formulação de respostas. Entretanto, esse mesmo grupo apresentou maior concordância com a afirmação de que teve dificuldade para compreender os problemas identificados (Q6). Entretanto, embora os participantes do GE tenham percebido maior facilidade na execução da tarefa (Q3 e Q4) e maior confiança em suas respostas (Q7 e Q8), esses ganhos não se refletiram em melhor desempenho na identificação de problemas de usabilidade.

5.2.2. Percepções específicas sobre o uso da IAGen

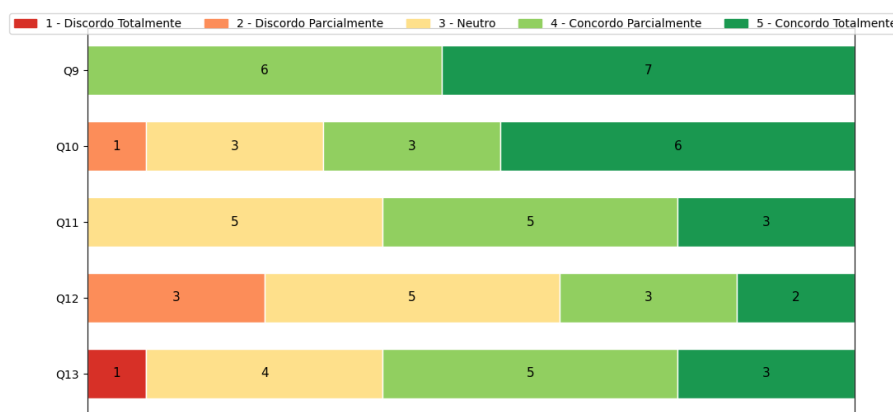
Considerando o contraste observado entre as percepções dos participantes e os resultados de desempenho, foram incluídas questões específicas para aprofundar a compreensão sobre o uso da IAGen durante a atividade. Para isso, os participantes do GE responderam a questões em escala Likert de cinco pontos e questões abertas relacionadas às contribuições, limitações e dificuldades percebidas durante o uso da ferramenta. A Tabela 3 apresenta a descrição das questões fechadas aplicadas exclusivamente a esse grupo.

Os resultados mostram elevada concordância para as questões relacionadas ao apoio fornecido pela ferramenta na compreensão dos problemas de usabilidade (Q9) e

Tabela 3. Questões fechadas específicas sobre o uso da IAGen

Código	Descrição
Q9	A IA me ajudou a compreender melhor os problemas
Q10	A IA facilitou a identificação de problemas de usabilidade
Q11	A IA influenciou minhas decisões durante a atividade
Q12	Eu segui as respostas da IA sem questionar
Q13	Confiei nas respostas fornecidas pela IA

na realização da atividade (Q10). Esses achados indicam que os estudantes perceberam a IAGen como um recurso útil para apoiar a interpretação dos problemas identificados e auxiliar o processo de avaliação. A Figura 3 apresenta a distribuição das respostas para essas questões.

**Figura 3. Frequência das respostas sobre o uso da IAGen do grupo GE**

De forma complementar, as respostas da questão Q11 indicam que a ferramenta exerceu influência nas decisões tomadas pelos participantes durante a atividade. Embora essa influência tenha sido percebida predominantemente de forma positiva, ela evidencia que as respostas produzidas pela IA desempenharam papel relevante no processo de análise conduzido pelos estudantes.

Em relação ao uso crítico da ferramenta, observa-se um comportamento mais heterogêneo. Enquanto as respostas da Q13 indicam níveis moderados a elevados de confiança nas respostas produzidas pela IA, as respostas da Q12 mostram que a maioria dos participantes não percebeu ter seguido as recomendações da ferramenta de forma totalmente automática ou acrítica. Em conjunto, esses resultados indicam que os estudantes acreditavam estar utilizando a IA de maneira consciente e controlada, ainda que reconhecessem sua influência sobre suas decisões.

As respostas abertas reforçam essa percepção. Os participantes relataram utilizar a IAGen principalmente para recordar heurísticas, validar interpretações e apoiar a classificação dos problemas encontrados. Por exemplo, P13 afirmou que a ferramenta o auxiliou a "lembrar as heurísticas", enquanto P14 destacou seu uso para "validar a relação das heurísticas com os problemas identificados". Os relatos também sugerem que a tomada de decisão permaneceu baseada principalmente na análise da interface, sendo a IA utilizada como mecanismo complementar de verificação. Conforme relatado por P16, os problemas eram inicialmente identificados e posteriormente associados às heurísticas

correspondentes, com apoio da ferramenta.

Em conjunto, os resultados indicam que os estudantes perceberam a IAGen como um recurso útil para apoiar a realização da atividade, especialmente na compreensão e aplicação das heurísticas de usabilidade. Entretanto, a combinação entre elevada confiança, influência percebida da ferramenta e ausência de ganhos observáveis de desempenho indica que a percepção positiva sobre a IA não foi acompanhada por melhorias objetivas na tarefa, reforçando a necessidade de investigar como as respostas produzidas pela ferramenta foram efetivamente utilizadas durante a atividade.

5.3. Análise dos Logs e Comportamentos de Uso da IA

A análise dos logs de interação do GE, disponibilizados de forma anonimizada nos artefatos suplementares do estudo, foi realizada em conjunto com o gabarito de referência utilizado na avaliação. Os registros permitiram identificar evidências de erros associados ao uso da IA, incluindo 2 casos de alucinação, 3 casos de erro induzido pela IA e 1 caso de sicofantia. Os casos identificados podem ser rastreados diretamente nos logs e nas planilhas de análise disponibilizadas no material suplementar.

Os resultados indicam que tais problemas não foram percebidos nem relatados pelos estudantes no pós-questionário. Em nenhuma das respostas abertas os participantes mencionaram ter identificado alucinações, informações incorretas ou problemas inventados pela ferramenta. Essa discrepância entre os registros das interações e os relatos dos participantes sugere limitações na capacidade de reconhecer respostas inadequadas produzidas pela IA.

A análise dos padrões de uso mostrou que apenas 2 dos 12 estudantes (16,67%) utilizaram a IA de forma ativa, empregando-a para validar hipóteses próprias, enquanto os demais 10 participantes (83,33%) adotaram um uso predominantemente passivo, solicitando respostas prontas ou explicações sem validação crítica. Os logs evidenciaram que os casos de alucinação e erro induzido ocorreram justamente em situações nas quais os participantes reproduziram respostas fornecidas pela IA sem questionamento adicional. Em um dos casos documentados, o estudante transcreveu problemas inexistentes sugeridos pela ferramenta. Em outro, a IA reforçou uma classificação incorreta inicialmente proposta pelo próprio participante, caracterizando um episódio de sicofantia.

6. Discussão e Implicações Educacionais

Os resultados mostram que a utilização da IAGen durante a atividade de Avaliação Heurística não resultou em melhorias significativas no desempenho dos estudantes. Embora os participantes do GE tenham relatado maior facilidade para executar a tarefa e maior confiança em suas respostas, essa percepção não foi acompanhada por um aumento na capacidade de identificar corretamente problemas de usabilidade. Esse resultado indica que o apoio fornecido pela IA pode gerar uma sensação de segurança que nem sempre corresponde à qualidade efetiva das análises produzidas.

A análise dos logs permitiu compreender melhor esse fenômeno, a maioria dos estudantes utilizou a ferramenta de forma predominantemente passiva, recorrendo às respostas geradas pela IA sem realizar verificações adicionais. Nesse contexto, foram observados casos de alucinação, erros induzidos pela IA e um episódio de sicofantia, nos quais informações incorretas foram aceitas ou reforçadas durante a atividade. O fato de

tais ocorrências não terem sido identificadas pelos participantes revela dificuldades na avaliação crítica das respostas produzidas pela ferramenta e aponta para a presença de processos de delegação cognitiva, nos quais parte do raciocínio analítico é transferida para o sistema.

Esses resultados reforçam que a integração da IAGen em contextos educacionais deve ser acompanhada por estratégias pedagógicas voltadas ao desenvolvimento do pensamento crítico e da literacia em IA. Mais do que fornecer acesso à tecnologia, torna-se necessário ensinar os estudantes a questionar, validar e justificar as informações obtidas, especialmente em atividades que exigem interpretação, análise e tomada de decisão. Nesse cenário, o papel do professor permanece essencial como mediador do processo de aprendizagem, promovendo práticas que estimulem a reflexão e reduzam a aceitação acrítica das respostas geradas por sistemas de IA [Favero et al. 2025].

7. Considerações Finais

Este trabalho investigou o uso de IAGen em atividades de Avaliação Heurística, analisando seus efeitos no desempenho, na percepção de aprendizagem e nos padrões de interação dos estudantes. Para isso, foi conduzido um experimento controlado com 24 estudantes de Engenharia de Software, distribuídos entre um Grupo Controle (sem IA) e um Grupo Experimental (com IA).

Em relação à QP1, os resultados não evidenciaram diferenças estatisticamente significativas entre estudantes que utilizaram IAGen e aqueles que realizaram a atividade sem seu auxílio, embora tenha sido observada uma tendência de desempenho superior no Grupo Controle. Quanto à QP2, os participantes do Grupo Experimental relataram maior facilidade para realizar a atividade, maior confiança em suas respostas e percepção positiva sobre o apoio fornecido pela ferramenta. Entretanto, essas percepções não foram acompanhadas por ganhos equivalentes de desempenho. Por fim, em resposta à QP3, a análise dos logs revelou predominância de uso passivo da ferramenta, ocorrência de delegação cognitiva, reprodução acrítica de respostas e dificuldades na identificação de erros produzidos pela IA, também foram observados episódios de alucinação e sicofantia que não foram percebidos ou reportados pelos participantes. Em conjunto, esses resultados indicam limitações no uso crítico da ferramenta e sugerem que os estudantes podem ter superestimado sua capacidade de avaliar a qualidade das respostas produzidas pela IA.

Os resultados deste estudo devem ser interpretados considerando algumas limitações, a pesquisa foi conduzida em uma única disciplina e instituição, envolvendo uma amostra relativamente pequena de estudantes. Além disso, a atividade foi realizada em um contexto específico de aprendizagem de Avaliação Heurística, o que pode limitar a generalização dos resultados para outros conteúdos, perfis de estudantes ou cenários educacionais. Também deve ser considerado que os participantes possuíam diferentes níveis de familiaridade com ferramentas de IA, fator que pode ter influenciado seus padrões de interação durante a atividade.

Como trabalhos futuros, pretende-se investigar estratégias de mediação pedagógica que promovam o uso crítico da IAGen, estimulando a validação, a argumentação e a reflexão sobre suas respostas. Também se prevê ampliar o estudo para outros contextos educacionais e avaliar intervenções que reduzam a delegação cognitiva e favoreçam interações mais ativas e reflexivas com sistemas de IA.

Disponibilidade de Artefato

Os artefatos completos usados para realização deste estudo estão disponíveis em: <https://doi.org/10.6084/m9.figshare.32519361>

Declaração sobre o uso da Inteligência Artificial

Os autores utilizaram o ChatGPT (versão GPT-5.5) como ferramenta de apoio na organização textual, revisão gramatical, aprimoramento da clareza da escrita e revisão estrutural do artigo. Todo o conteúdo produzido com auxílio da ferramenta foi revisado, validado e editado pelos autores, que se responsabilizam integralmente pelas informações, interpretações e conclusões apresentadas no trabalho.

Agradecimentos

Agradecemos a todos os participantes do estudo experimental e aos membros dos grupos de pesquisa USES e TechThinkers, pelo apoio, colaboração e incentivo durante o desenvolvimento deste trabalho. Agradecemos também ao Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM), especialmente ao Campus Manaus Zona Leste, pelo suporte institucional e pelas condições proporcionadas para a realização desta pesquisa.

Este trabalho contou com o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), por meio dos processos nº 314797/2023-8, 443934/2023-1 e 445029/2024-2; da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES), por meio do AUXPE-CAPES-PROEX, Código de Financiamento 001; e da Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM), por meio dos projetos PDPG-CAPES, POSGRAD 2025–2026 e POSGRAD 2026–2027.

Referências

- Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier.
- Bahroun, Z., Anane, C., Ahmed, V., and Zacca, A. (2023). Transforming education: A comprehensive review of generative artificial intelligence in educational settings through bibliometric and content analysis. *Sustainability*.
- Baidoo-Anu, D. and Ansah, L. O. (2023). Education in the era of generative artificial intelligence (ai): Understanding the potential benefits of chatgpt in promoting teaching and learning. *SSRN Electronic Journal*.
- Barbosa, S. D. J., Silva, B. d., Silveira, M. S., Gasparini, I., Darin, T., and Barbosa, G. D. J. (2021). Interação humano-computador e experiência do usuário. *Auto publicação*.
- Barke, S., James, M. B., and Polikarpova, N. (2023). Grounded copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages*, 7(OOPSLA1):85–111.
- Bauschard, S. and Quidwai, S. (2024). From insight to implementation: How to create your ai school guidance. *Available at SSRN 4784207*.

- Becker, B. A., Craig, M., Denny, P., Keuning, H., Kiesler, N., Leinonen, J., Luxton-Reilly, A., Prather, J., and Quille, K. (2023). Generative ai in introductory programming. *Computer Science Curricula*, pages 438–439.
- Branco, B. C., Neto, A. M., Avelino, G., and Machado, V. (2025). Avaliando o papel do gpt-4o na aprendizagem de engenharia de software: Um relato de experiência. In *Anais do XXXIX Simpósio Brasileiro de Engenharia de Software*, pages 470–481, Porto Alegre, RS, Brasil. SBC.
- Chan, C. K. Y. and Hu, W. (2023). Students’ voices on generative ai: Perceptions, benefits, and challenges in higher education. *International journal of educational technology in higher education*, 20(1):43.
- De Cremer, D., Bianzino, N. M., and Falk, B. (2023). How generative ai could disrupt creative work. *Harvard Business Review*, 13:13.
- Dibia, V., Fourney, A., Bansal, G., Poursabzi-Sangdeh, F., Liu, H., and Amershi, S. (2023). Aligning offline metrics and human judgments of value for code generation models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8516–8528.
- Favero, L., Pérez-Ortiz, J.-A., Käser, T., and Oliver, N. (2025). Do ai tutors empower or enslave learners? toward a critical use of ai in education. *arXiv preprint arXiv:2507.06878*.
- Gerlich, M. (2025). Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6.
- Guerino, G., Rodrigues, L., Capeleti, B., Mello, R., Freire, A., and Zaina, L. A. M. (2025). Can gpt-4o evaluate usability like human experts? a comparative study on issue identification in heuristic evaluation. pages 381–402.
- Hong, H., Vate-U-Lan, P., and Viriyavejakul, C. (2025). Cognitive offload instruction with generative ai: A quasi-experimental study on critical thinking gains in english writing. In *Forum for Linguistic Studies*, volume 7, pages 325–334.
- Lazar, J., Feng, J. H., and Hochheiser, H. (2017). *Research methods in human-computer interaction*. Morgan Kaufmann.
- Mittal, U., Sai, S., Chamola, V., and Sangwan, D. (2024). A comprehensive review on generative ai for education. *Ieee Access*, 12:142733–142759.
- Natella, R., Winter, S., Cotroneo, D., and Suri, N. (2018). Analyzing the effects of bugs on software interfaces. *IEEE Transactions on Software Engineering*, 46(3):280–301.
- Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 152–158.
- Nielsen, J. (1994b). Usability inspection methods. In *Conference companion on Human factors in computing systems*, pages 413–414.
- Ravšelj, D., Keržič, D., Tomaževič, N., Umek, L., Brezovar, N., Iahad, N. A., Abdulla, A. A., Akopyan, A., Segura, M. W. A., AlHumaid, J., et al. (2025). Higher education students’ perceptions of chatgpt: A global study of early reactions. *PLoS One*, 20(2):e0315011.

- Rocha, H. and Baranauskas, M. C. C. (2003). Design e avaliação de interfaces humano-computador. *Campinas: Unicamp*.
- Sajja, R., Sermet, Y., Fodale, B., and Demir, I. (2025). Evaluating ai-powered learning assistants in engineering higher education: Student engagement, ethical challenges, and policy implications. *ArXiv*, abs/2506.05699.
- Silvestre, A. S. S., de Moura Amaral, E. L., Holanda, M. E., and Canedo, E. D. (2023). Students' perception about chatgpt's impact on their academic education. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1260–1270. SBC.
- Sobrino-Duque, R., Martínez-Rajo, N., Carrillo-de Gea, J. M., López-Jiménez, J. J., Nicolás, J., and Fernández-Alemán, J. L. (2022). Evaluating a gamification proposal for learning usability heuristics: Heureka. *International Journal of Human-Computer Studies*, 161:102774.
- Sousa, H. and Cruz, D. (2024). Capacitando educadores com ia generativa: Implicações na educação. In *Anais do XXXV Simpósio Brasileiro de Informática na Educação*, pages 1931–1941, Porto Alegre, RS, Brasil. SBC.
- Triberti, S., Di Fuccio, R., Scuotto, C., Marsico, E., and Limone, P. (2024). “better than my professor?” how to develop artificial intelligence tools for higher education. *Frontiers in Artificial Intelligence*, 7:1329605.
- Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P. S., and Wen, Q. (2025). Large language models for education: A survey and outlook. *IEEE Signal Processing Magazine*, 42(6):51–63.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer, Berlin, Germany.
- Xia, Q., Weng, X., Fan, O., Lin, T.-J., and Chiu, T. (2024). A scoping review on how generative artificial intelligence transforms assessment in higher education. *International Journal of Educational Technology in Higher Education*, 21:1–22.
- 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.
- Zhan, Z., Wu, Q., Lin, Z., and Cai, J. (2021). Smart classroom environments affect teacher-student interaction: Evidence from a behavioural sequence analysis. *Australasian Journal of Educational Technology*, 37(2):96–109.
- Zhong, R., McDonald, D. W., and Hsieh, G. (2025). Synthetic heuristic evaluation: A comparison between ai- and human-powered usability evaluation. *ArXiv*, abs/2507.02306.