Avaliação do Uso de LLMs na Geração de Casos de Teste a Partir de *User Stories*: Um Estudo Experimental em Contexto Educacional com Análise de *Test Smells*

Juliana B. Lima[1], **Márcia Sampaio Lima**[1]

[1]Escola Superior de Tecnologia – Universidade do Estado do Amazonas (UEA)
Caixa Postal 15.064 – 69.020-120 – Manaus – AM – Brasil

{jbl.snf23, msllima}@uea.edu.br

Abstract. Introduction & objective*: This experimental study investigates the educational use of LLMs in software testing, focusing on test cases (TCs) generation from user stories and quality analysis through test smells.* ***Steps****: Involving 25 students, the study compared manual and ChatGPT-assisted TCs generation, followed by test smell identification.* ***Results:*** *Most ChatGPT cases were considered useful (83%), though only 31% were novel. Manual cases showed a higher prevalence of specific test smells, such as Generic Expected Result (44% vs. 0%). Participants perceived the experience positively for learning while noting concerns about LLM dependency. The study provides empirical evidence on LLMs in testing education.*
Keywords: *Test Cases, Generative IA, Computing Education*

Resumo. Introdução & objetivo: *Este estudo investiga o uso educacional de LLMs em testes de software, focando na geração de casos de teste (CTs) a partir de User Stories e análise de qualidade via test smells.* ***Etapas:*** *Envolvendo 25 estudantes, o estudo comparou a geração de CTs manualmente e a assistida por ChatGPT, seguida da identificação de test smells.* ***Resultados:*** *A maioria dos CTs gerados pelo ChatGPT foi considerada útil (83%), mas apenas 31% foram considerados novos na perspectiva dos estudantes. Casos manuais apresentaram maior prevalência de smells como "Resultado Esperado Genérico" (44% vs. 0%). Os participantes avaliaram a experiência positivamente para aprendizagem e apontaram preocupações com a dependência da ferramenta. O estudo apresenta evidências experimentais sobre LLMs na educação em teste de software.*
Palavras-chave: *Casos de Teste, IA Generativa, Ensino em computação*

1. Introdução

A integração da Inteligência Artificial Generativa, particularmente dos Modelos de Linguagem de Grande Escala (LLMs), no ensino de Engenharia de Software (ES) tem emergido como uma área de interesse na Educação em Computação [Rodrigues et al. 2024]. No ensino de testes de software, a geração de casos de teste a partir de requisitos exige compreensão do domínio, raciocínio lógico e conhecimento técnico, tornando-se uma atividade adequada para práticas pedagógicas que combinam produção, comparação e análise crítica de artefatos. Estudos recentes demonstram o potencial dos LLMs para

apoiar essa tarefa [Manzoni et al. 2024]. Pesquisas sobre ChatGPT no ensino de testes também indicam oportunidades pedagógicas, mas apontam riscos relacionados à dependência da ferramenta e à compreensão conceitual dos estudantes [Jalil et al. 2023, Mezzaro et al. 2024]. Porém, ainda permanecem lacunas sobre a qualidade dos artefatos gerados e sobre como esse apoio pode contribuir para a aprendizagem de estudantes.

Neste cenário, apresenta-se um relato de experiência com caráter experimental exploratório, referente à implementação de uma atividade em um curso real de Engenharia de Software. A experiência envolveu a comparação entre casos de teste produzidos manualmente e aqueles gerados com apoio do ChatGPT, permitindo analisar decisões metodológicas, desafios observados e aprendizagens pedagógicas emergentes dessa prática.

Uma evidência indicativa da qualidade dos casos de teste gerados é a presença ou ausência de *test smells*. De acordo com Aranda et al. [Aranda et al. 2024], *test smells* são inconsistências em artefatos de teste que prejudicam a legibilidade, manutenção e confiabilidade, como por exemplo, ambiguidade na descrição dos passos (*Ambiguous Language*), uso de vocabulário inconsistente (*Inconsistent Vocabulary*) ou resultados esperados genéricos (*Generic Expected Result*) [Aranda et al. 2024, Misu et al. 2025]. A presença de tais problemas nos casos de teste pode comprometer a eficácia dos testes e introduzir custos adicionais de manutenção do sistema [Aranda et al. 2024]. Diante deste cenário, a identificação e a correção desses problemas constituem habilidades importantes para estudantes de ES, contribuindo para o desenvolvimento do pensamento crítico sobre a qualidade do software.

O estudo investigou três dimensões principais: (i) a geração de casos de teste a partir de *user stories* (USs) com e sem apoio do ChatGPT; (ii) a análise comparativa da qualidade dos casos por meio da ocorrência de *test smells*; e (iii) a percepção dos estudantes quanto ao impacto pedagógico da abordagem. A partir dessas dimensões, foram formuladas as seguintes questões de pesquisa:

- **QP1:** Como os casos de teste gerados por LLMs se comparam com os produzidos manualmente em termos de utilidade percebida e de novidade?
- **QP2:** Quais são as diferenças na prevalência de *test smells* entre casos de teste manuais e os gerados por LLMs?
- **QP3:** Qual o impacto percebido da utilização de LLMs na experiência de aprendizagem de testes de software?

Para responder a essas questões, foi conduzido um experimento com estudantes de graduação em Computação, seguindo uma metodologia de estudo mista. **A principal contribuição deste trabalho é dupla:** primeiro, fornece evidências experimentais sobre as capacidades e limitações do ChatGPT em atividades educacionais de geração de casos de teste de software; segundo, propõe uma abordagem pedagógica que integra a geração assistida por IA com desenvolvimento de pensamento crítico através da análise de qualidade. Os resultados têm implicações tanto para educadores que buscam incorporar ferramentas de IA em seus currículos quanto para pesquisadores interessados em métodos experimentais para avaliação de tecnologias generativas em contextos educacionais.

2. Fundamentação Teórica e Trabalhos Relacionados

O uso da **IA Generativa em Contexto Educacional** requer um equilíbrio entre a potencialização do aprendizado e a cautela pedagógica [Crompton 2023,

Pitts et al. 2025]. Experimentos educacionais com IA no contexto de ES, como os de Queiroz e Lima [Queiroz and Lima 2025] e Rodrigues, Manzoni e Rocha [Rodrigues et al. 2024], evidenciam ganhos pedagógicos ao associar LLMs à aprendizagem prática, fornecendo referências metodológicas para a organização de artefatos e a condução de experimentos. Esses estudos enfatizam a importância da mediação pedagógica para garantir que o uso de IA complemente, e não substitua, o desenvolvimento de habilidades fundamentais.

Já **a qualidade das *user stories* (USs)** constitui um fator crítico para o sucesso do desenvolvimento ágil e a subsequente geração de CTs. O *framework* INVEST (*Independent, Negotiable, Valuable, Estimable, Small, Testable*), proposto por Cohn [Cohn 2004], fornece critérios bem estabelecidos para a avaliação da qualidade desses artefatos. Estudos experimentais, como o de Kuhail [Kuhail et al. 2022], validam a correlação entre a aderência aos critérios INVEST e a qualidade percebida das USs. Pesquisas recentes investigam a capacidade dos LLMs em gerar USs de qualidade. Dutta e Bhowmick [Dutta and Bhowmick 2025] demonstraram que, embora o ChatGPT produza histórias bem estruturadas linguisticamente, ainda apresenta limitações relacionadas à ambiguidade e redundância, reforçando a necessidade de supervisão humana. Esta pesquisa utiliza exemplos validados da Mountain Goat Software [Cohn 2024] como referência de qualidade para a *user story* utilizada no experimento.

***Test smells* e refatoração em linguagem natural** representam sintomas de problemas de qualidade em artefatos de teste que prejudicam a legibilidade, a manutenção e a confiabilidade [Misu et al. 2025]. Aranda et al. [Aranda et al. 2024] propõem um catálogo de sete transformações para eliminar esses problemas em testes manuais escritos em linguagem natural, servindo como referência teórica para este trabalho. O catálogo inclui sete *smells*: *Ambiguous Language, Inconsistent Vocabulary, Redundancy, Long Test, Irrelevant Inputs, Generic Expected Result,* e *Inadequate Granularity*. Pesquisas emergentes avaliam a capacidade dos LLMs em reconhecer e corrigir esses defeitos de forma autônoma [Santana Jr. et al. 2025, Ouédraogo et al. 2024], abrindo novas perspectivas para garantia de qualidade em testes, especialmente em contextos educacionais onde o desenvolvimento de habilidades de análise crítica é essencial.

O uso de LLMs na geração de casos de teste tem como base teórica o uso de técnicas de Processamento de Linguagem Natural (NLP) para extrair informações de requisitos ágeis [Raharjana et al. 2021]. Avanços recentes demonstram progresso na geração automatizada de casos de teste a partir de descrições textuais. Alagarsamy et al. [Alagarsamy et al. 2025] propõem o *fine-tuning* de LLMs para transformação de requisitos em casos de teste, enquanto Li et al. [Li et al. 2025] oferecem uma avaliação do desempenho de LLMs em múltiplas tarefas de teste de software. No contexto educacional, Manzoni, Rodrigues e Rocha [Manzoni et al. 2024] exploraram a geração de testes baseados em USs com ChatGPT, identificando ganhos em produtividade, mas também limitações quanto à clareza e completude.

No ensino de testes de software, estudos recentes têm investigado LLMs como apoio à aprendizagem conceitual, à geração de artefatos de teste e à reflexão sobre qualidade [Jalil et al. 2023, Mezzaro et al. 2024, Haldar et al. 2025]. Esta pesquisa avança nessa direção ao incorporar a análise de *test smells* como mecanismo de avaliação de qualidade em uma experiência educacional, investigando a geração inicial dos CTs e a

capacidade dos estudantes de comparar, avaliar e discutir os artefatos produzidos.

3. Metodologia da Pesquisa

Com base no paradigma Goal Question Metric (GQM), o objetivo deste estudo foi analisar o uso do ChatGPT na geração de CTs a partir de USs, com o propósito de avaliar sua utilidade educacional, a novidade percebida dos CTs gerados, a ocorrência de *test smells* e a percepção de aprendizagem dos estudantes, sob o ponto de vista de docentes e pesquisadores de Educação em Computação, no contexto de uma atividade prática de Engenharia de Software em sala de aula.

Por se tratar de um relato de experiência com análise descritiva, não foram definidos testes de hipótese inferenciais. Em vez disso, o estudo foi orientado por expectativas analíticas associadas às QPs: espera-se que os CTs gerados por LLMs sejam percebidos como úteis pelos estudantes; espera-se diferença na prevalência de *test smells* entre CTs manuais e CTs gerados por LLMs; e espera-se que a atividade contribua para a percepção de aprendizagem sobre geração e avaliação de CTs.

***Design* do estudo:** Este estudo adota uma abordagem mista, quantitativa e qualitativa, estruturada como intervenção pedagógica em ambiente educacional, com análise descritiva dos dados produzidos pelos estudantes. A pesquisa foi estruturada em três encontros presenciais, durante a disciplina de Tópicos Especiais em Engenharia de Software da Universidade do Estado do Amazonas (UEA).

Participantes: Participaram do estudo 25 estudantes matriculados na disciplina de Tópicos Especiais em Engenharia de Software, vinculados aos cursos de Sistemas de Informação, Licenciatura em Computação e Engenharia da Computação. Todos completaram as etapas do experimento, desde a assinatura do TCLE até os questionários finais. A caracterização do grupo considerou curso, período/semestre, experiência prévia com testes de software, experiência profissional e conhecimento prévio sobre CTs e *test smells*, informações coletadas por questionário inicial e disponibilizadas no material suplementar.

Procedimento experimental: O experimento foi conduzido em três aulas consecutivas de aproximadamente 90 minutos cada. Na Aula 1 (Preparação Conceitual), realizou-se uma exposição teórica sobre USs, CTs e os conceitos de *Behavior-Driven Development* (BDD), com exemplos práticos para nivelar o conhecimento dos participantes. Na Aula 2 (Fase 1 – Geração de CTs), os participantes elaboraram, inicialmente, cinco CTs manualmente a partir de uma *user story* predefinida (40 min). Em seguida, utilizaram um *prompt* padronizado no ChatGPT (versão gratuita) para gerar cinco novos CTs (30 min), avaliando cada um quanto à novidade, utilidade e adequação. O limite de cinco CTs foi adotado em razão do tempo disponível para execução da atividade em sala de aula e para manter a tarefa padronizada entre os participantes. Ao final, responderam a um questionário com escala Likert (1–5) e questões abertas sobre a experiência. Na Aula 3 (Fase 2 – Análise de *Test Smells*), foi realizada uma exposição teórica sobre *test smell*, baseada no catálogo de Aranda et al. [Aranda et al. 2024]. Em seguida, os participantes identificaram e corrigiram *test smells* em dois CTs pré-selecionados (um manual e um gerado pelo ChatGPT), utilizando o conjunto de sete *smells* (45 min). Ao final, os participantes preencheram um questionário avaliativo sobre a contribuição da atividade para a compreensão do tema.

Instrumentos de coleta de dados: Foram utilizados múltiplos instrumentos de

coleta ao longo do experimento. No início da Aula 1, os participantes assinaram o TCLE e responderam ao questionário de caracterização, que incluía informações sobre período/semestre, contato prévio com testes de software, experiência profissional e autoavaliação de conhecimento prévio sobre CTs e *test smells*. Durante a Aula 2 (Fase 1), foram coletadas 125 avaliações dos CTs gerados pelo ChatGPT. A novidade foi entendida como a percepção de que o CT apresentava uma ideia não elaborada previamente pelo estudante, enquanto a utilidade foi entendida como a percepção de que o CT poderia contribuir para verificar o comportamento esperado da US. As respostas sobre novidade e utilidade foram categorizadas a partir das alternativas do formulário. Além disso, foram coletadas respostas em escala Likert de 5 pontos sobre a contribuição percebida da experiência para a compreensão da atividade de geração de CTs. Na Aula 3 (Fase 2), foram analisados 50 CTs (25 manuais e 25 do ChatGPT) para as sete categorias de *test smells* e coletados dados sobre clareza percebida, confiança e contribuição da experiência para a compreensão de *test smells* (escala Likert). Dados qualitativos incluíram respostas abertas, pontos positivos/negativos e percepções sobre a experiência em ambas as fases.

Análise de dados: A análise empregou métodos mistos para integrar perspectivas quantitativas e qualitativas. Para os dados quantitativos, foi realizada análise estatística descritiva, incluindo cálculo de proporções e frequências de respostas nas escalas Likert, considerando a natureza ordinal dos dados. Não foram aplicados testes estatísticos inferenciais, pois o foco do trabalho é descrever e discutir uma experiência educacional em sala de aula, e não estabelecer generalizações estatísticas a partir da amostra. Foram realizadas comparações entre CTs manuais (gerados sem apoio do ChatGPT) e gerados com apoio do ChatGPT quanto à prevalência de diferentes categorias de *test smells*. Para os dados qualitativos, foi realizada análise de conteúdo temática, seguindo abordagem indutiva [Braun and Clarke 2006]. As respostas abertas foram codificadas para identificar temas recorrentes, com categorização em pontos positivos e negativos. Para garantir a consistência na análise de *test smells*, foi realizado um processo de correção e revisão dos CTs gerados na Fase 2. Esse processo foi conduzido pelos autores deste artigo, com discussão de casos duvidosos para estabelecer consenso nas classificações.

Disponibilidade de materiais: Os *scripts* de atividades utilizados no experimento, os resultados brutos, as análises completas realizadas com *Google Colab* e *Python*, além de artefatos complementares, estão disponíveis publicamente no repositório Figshare [Lima and Lima 2025].

4. Resultados

Os resultados são apresentados conforme as três questões de pesquisa, mantendo correspondência entre os instrumentos de coleta, as métricas analisadas e as evidências produzidas em cada fase do estudo. **Resultados da Fase 1: Geração de CTs.** A avaliação de 125 CTs gerados pelo ChatGPT revelou percepções predominantemente positivas entre os participantes. Como mostra a Tabela 1, aproximadamente um terço (31%) dos CTs foi considerado novo (não elaborado manualmente), enquanto 83% foram considerados úteis. Cerca de um quarto dos CTs (24%) eram simultaneamente novos e úteis. A Figura 1 mostra que os CTs considerados novos também apresentaram alta taxa de utilidade, com 77% deles avaliados como úteis. Os CTs não considerados novos, isto é, aqueles já elaborados manualmente pelos estudantes, apresentaram utilidade ainda maior, com 88% de avaliações positivas. Os CTs parcialmente similares aos manuais também foram con-

siderados úteis em sua maioria. A Figura 2 indica variação na novidade percebida entre os CTs gerados, com o quinto caso apresentando maior proporção de casos novos.

Tabela 1. Síntese dos Resultados da Fase 1 (n = 125 casos)

Métrica	Proporção (%)
CTs não elaborados manualmente (novidade)	31%
CTs considerados úteis	83%
CTs novos e úteis	24%
CTs similares aos manuais (reutilizados total ou parcialmente)	69%

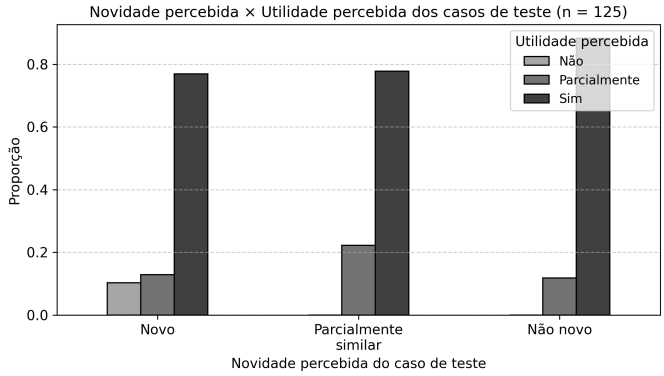

Figura 1. Relação entre novidade e utilidade percebidas dos CTs - ChatGPT.

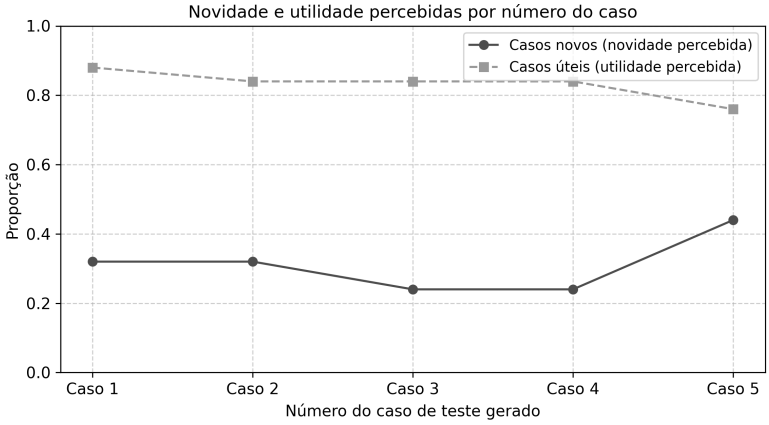

Figura 2. Proporção de casos considerados novos e úteis por CT - ChatGPT

Percepção qualitativa da experiência (Fase 1) revelou 29 menções positivas e 28 negativas. Entre os aspectos positivos, destacaram-se agilidade e praticidade (P4: *"rápido e prático"*), maior formalidade e objetividade dos CTs (P9: *"mais formal e objetivo"*), apoio ao aprendizado (P2: *"me deu uma visão melhor do que são CTs"*) e identificação de cenários não considerados (P18: *"indicou um caso que eu não havia percebido"*). Entre os pontos negativos foram mencionadas inconsistências e falta de especificidade (P12: *"falta de especificidade no teste"*), risco de dependência da ferramenta (P19: *"sensação de conforto e dependência"*) e redução do esforço cognitivo (P10: *"não tive que pensar*

sobre"). Ainda assim, 88% dos participantes atribuíram notas 4 ou 5 à contribuição da experiência para a compreensão da atividade, indicando avaliação positiva.

Resultados da Fase 2: Análise de *Test Smells*. A Tabela 2 apresenta a prevalência de cada tipo de *test smell* nos CTs analisados, comparando aqueles gerados manualmente com os gerados pelo ChatGPT. Os resultados indicam que os casos manuais apresentaram maior ocorrência nas sete categorias de *test smells* analisadas, especialmente "Resultado Esperado Genérico" (44% dos casos manuais vs. 0% dos casos do ChatGPT) e "Ambiguidade" (48% vs. 36%). Para o cálculo dessas proporções, considerou-se, para cada grupo (manual e ChatGPT), o número de casos que continham pelo menos uma instância de cada tipo de *test smell*, dividido pelo total de casos analisados naquele grupo (25 casos de cada tipo). Por exemplo, dos 25 casos manuais, 11 apresentaram o *smell* "Resultado Esperado Genérico", resultando em 44%, enquanto nenhum dos 25 casos do ChatGPT apresentou esse *smell* (0%).

Tabela 2. Prevalência de Test Smells por Origem do Caso

Tipo de Test Smell	Manual (%)	ChatGPT (%)
Ambiguidade	48%	36%
Vocabulário Inconsistente	16%	8%
Redundância	20%	8%
Teste Longo	20%	4%
Entradas Irrelevantes	8%	4%
Resultado Esperado Genérico	44%	0%
Granularidade Inadequada	36%	20%

Tabela 3. Qualidade da Identificação e Correção dos *Smells* por Origem

Métrica	Manual (%)	ChatGPT (%)
Smells identificados estavam presentes	91%	58%
Havia *smells* adicionais não identificados	48%	8%
Smells identificados foram corrigidos	89%	58%

A Figura 3 ilustra essa comparação por meio de um gráfico de barras agrupadas, permitindo visualizar as diferenças na ocorrência de cada tipo de *test smell* entre os casos de teste manuais e os gerados pelo ChatGPT. Observa-se que os casos manuais (representados por barras em tonalidade mais clara) apresentaram prevalência superior em todas as categorias de *smells*. As maiores diferenças entre os grupos ocorreram em "Resultado Esperado Genérico" (44% vs. 0%), "Teste Longo" (20% vs. 4%) e "Granularidade Inadequada" (36% vs. 20%). Além disso, "Ambiguidade" apresentou a maior prevalência geral nos dois grupos (48% nos casos manuais e 36% nos casos gerados pelo ChatGPT). Por outro lado, ambos os grupos apresentaram baixa ocorrência de "Entradas Irrelevantes" (8% vs. 4%) e "Teste Longo" (20% vs. 4%).

As menores diferenças ocorreram em "Entradas Irrelevantes" (8% vs. 4%), "Vocabulário Inconsistente" (16% vs. 8%) e "Redundância" (20% vs. 8%). A Tabela 3 revela que, em CTs manuais, 91% dos smells identificados pelos participantes estavam realmente presentes nos casos (vs. 58% para ChatGPT), e a taxa de correção bem-sucedida

dos *smells* identificados foi maior (89% vs. 58%). Essas últimas métricas avaliam a acurácia da identificação e eficácia da correção realizadas pelos estudantes.

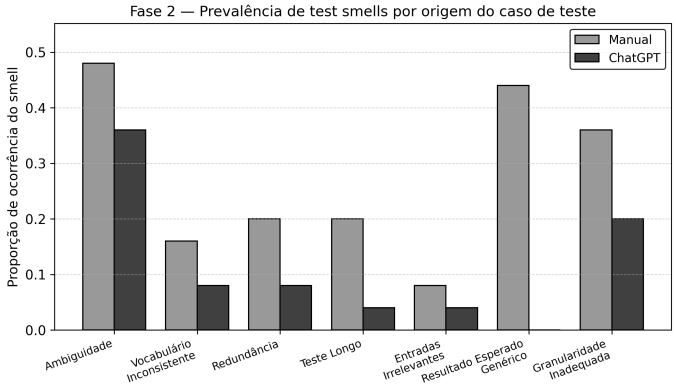

Figura 3. Prevalência de *test smells* em CTs manuais e gerados pelo ChatGPT.

Conhecimento e clareza sobre *test smells*: Antes da aula, 68% dos participantes declararam não conhecer o conceito de *test smells*. Após a exposição teórica e a atividade prática, 88% avaliaram o conceito como claro ou muito claro. Quanto à contribuição da Fase 2 para a compreensão sobre criação e qualidade de CTs, 88% atribuíram notas 4 ou 5. Após a experiência, 80% relataram sentir-se mais confiantes para criar CTs, enquanto 20% indicaram confiança parcial.

Percepção qualitativa da experiência (Fase 2) registrou 19 menções positivas e 16 negativas. Entre os aspectos positivos, destacaram-se a aplicação prática dos conceitos (P6: *"A prática ajudou muito na construção desse conhecimento"*), o desenvolvimento da análise crítica (P18: *"tenho uma ideia melhor sobre como fazer casos de teste e avaliá-los"*) e a introdução a um tema relevante (P16: *"conceitos dos smells que não conhecia e parecem importantes"*). Como pontos negativos foram mencionados a dificuldade de identificação dos *smells* (P17: *"foi difícil encontrar os smells"*), o tempo limitado para a atividade (P4: *"Mais tempo para absorver os tipos de smells"*), a subjetividade de alguns critérios (P7: *"dificuldade em achar um meio termo"*) e a necessidade de maior aprofundamento (P24: *"o conteúdo poderia ser mais profundo"*).

Em síntese, respondendo às questões de pesquisa formuladas neste estudo: (QP1) os casos de teste gerados por LLMs comparam-se aos manuais com alta utilidade percebida (83%) mas limitada novidade (31%); (QP2) casos manuais apresentaram maior prevalência em todas as sete categorias de *test smells* analisadas, com maior diferença em "Resultado Esperado Genérico" (44% vs. 0%) e diferenças relevantes em "Teste Longo" (20% vs. 4%) e "Granularidade Inadequada" (36% vs. 20%). Além disso, "Ambiguidade" apresentou alta prevalência nos dois grupos (48% vs. 36%), embora a identificação e correção de *smells* tenham sido mais precisas em casos manuais (91% de acurácia vs. 58% para casos do ChatGPT); e (QP3) o impacto percebido na aprendizagem foi predominantemente positivo, com 88% dos participantes avaliando a experiência como contributiva para a compreensão da geração de CTs e da análise de *test smells*, 80% relatando maior confiança para criar CTs, e percepções qualitativas destacando benefícios como agilidade e apoio ao aprendizado, embora com preocupações sobre dependência da ferra-

menta e redução do esforço cognitivo.

5. Discussão

Os resultados deste estudo indicam que o ChatGPT apresenta alta utilidade percebida (83%) na geração de CTs, corroborando pesquisas que apontam ganhos em produtividade e qualidade [Manzoni et al. 2024, Alagarsamy et al. 2025]. Contudo, a baixa proporção de casos considerados novos (31%) sugere que a ferramenta tende a gerar casos semelhantes aos produzidos manualmente pelos estudantes, reforçando práticas já consolidadas mais do que ampliando criativamente o espaço de testes. Assim, seu valor pedagógico parece residir principalmente na validação e no refinamento das ideias dos estudantes.

A maior prevalência de *test smells* em casos manuais, especialmente "Resultado Esperado Genérico" (44% vs. 0%) e "Ambiguidade" (48% vs. 36%), indica que o ChatGPT pode produzir resultados esperados mais específicos e descrições menos ambíguas. Entretanto, os estudantes apresentaram menor acurácia na identificação de *smells* em casos gerados automaticamente (58% vs. 91%), sugerindo que a estrutura aparentemente mais formal desses textos pode dificultar a análise crítica.

As percepções qualitativas evidenciam uma postura reflexiva dos participantes: embora reconheçam benefícios como agilidade e apoio ao aprendizado, apontam riscos de dependência e redução do esforço cognitivo, em consonância com discussões da literatura sobre o uso equilibrado de IA generativa na educação [Crompton 2023, Pitts et al. 2025]. Em conjunto, os resultados reforçam a necessidade de mediação pedagógica no uso de LLMs, de modo que a ferramenta complemente, e não substitua, o desenvolvimento de habilidades analíticas fundamentais. A contribuição do estudo é principalmente educacional e incremental: a experiência mostra como a geração assistida por LLMs pode ser usada como ponto de partida para comparação, discussão e análise crítica de CTs em uma disciplina de Engenharia de Software.

6. Limitações

Este estudo apresenta limitações que devem ser consideradas na interpretação dos resultados. A participação voluntária de 25 estudantes de uma única instituição pode introduzir viés de seleção e limitar a representatividade para contextos mais amplos, acadêmicos ou industriais. O perfil específico dos participantes, incluindo curso, período e experiência prévia com testes, também pode ter influenciado a produção e a avaliação dos CTs. Além disso, o efeito de aprendizagem entre as fases e o tempo limitado das atividades podem ter influenciado o desempenho na etapa subsequente.

Quanto à generalização, o experimento foi conduzido com uma única *user story* em domínio específico, o que restringe a extensão dos resultados para outros tipos de requisitos. O uso de um *prompt* padronizado permitiu controlar a atividade, mas também limita a análise de outras formas de interação com LLMs. O uso exclusivo da versão gratuita do ChatGPT também limita a comparação com outros modelos ou versões mais avançadas.

Na operacionalização dos construtos, medidas como utilidade percebida e novidade percebida baseiam-se em avaliações subjetivas, podendo conter vieses. A definição de *test smells* segue um catálogo específico [Aranda et al. 2024], que pode não abranger todas as dimensões de qualidade em testes escritos em linguagem natural. Por fim,

a análise qualitativa, embora relevante, está sujeita à interpretação dos pesquisadores. Também é possível que a maior formalidade dos CTs gerados por LLMs tenha produzido uma percepção de correção aparente, dificultando a identificação de problemas pelos estudantes. Essa limitação deve ser investigada em estudos posteriores.

7. Considerações Éticas

Esta pesquisa foi conduzida em conformidade com os princípios éticos acadêmicos e as diretrizes institucionais para pesquisas envolvendo seres humanos. Todos os participantes foram devidamente informados sobre os objetivos, procedimentos e finalidades do estudo através do TCLE, que explicitava: **Voluntariedade** (participação inteiramente voluntária, com garantia de desistência a qualquer momento sem penalidades acadêmicas); **Confidencialidade** (anonimização completa dos dados pessoais, com identificação apenas através de códigos P1 a P25); **Uso dos Dados** (utilização exclusiva para fins de pesquisa acadêmica); **Aspectos educacionais** (caráter educacional da pesquisa e alinhamento com os objetivos da disciplina); e **Transparência** (divulgação completa dos procedimentos). O estudo foi desenvolvido no âmbito do Programa de Apoio à Iniciação Científica (PAIC) da UEA. O estudo foi conduzido como atividade educacional de baixo risco no contexto regular da disciplina, sem coleta de dados sensíveis e sem identificação dos participantes. Ainda assim, foram adotados TCLE, participação voluntária, anonimização dos dados e possibilidade de desistência sem prejuízo acadêmico. Foram tomados cuidados para garantir conformidade com os termos de serviço do ChatGPT e políticas institucionais sobre uso de IA generativa.

8. Conclusão e Trabalhos Futuros

Este estudo investigou o uso de LLMs no ensino de testes de software, integrando geração assistida de CTs com análise de qualidade via *test smells*. Os resultados indicam que o ChatGPT pode ser uma ferramenta valiosa no contexto educacional quando utilizado como apoio à prática, à comparação e à reflexão crítica sobre CTs, e não apenas como gerador automático de respostas. No campo da Educação em Computação, a experiência reforça a importância de atividades mediadas por docentes para desenvolver autonomia, senso crítico e compreensão sobre qualidade em testes. As principais contribuições são: (1) evidências experimentais sobre utilidade e novidade de CTs gerados por LLMs; (2) análise comparativa da prevalência de *test smells*; (3) *insights* qualitativos sobre percepções dos estudantes; (4) uma abordagem pedagógica replicável; e (5) identificação de desafios na análise de CTs gerados automaticamente.

Para trabalhos futuros, sugerem-se estudos com amostras maiores e mais diversas; investigação de diferentes modelos de LLMs; expansão para múltiplos domínios; desenvolvimento de estratégias pedagógicas específicas para mediação do uso de LLMs; e investigação do impacto de diferentes formatos de prompt. A integração responsável de IA generativa na educação representa uma oportunidade e um desafio que requer abordagens cuidadosas, e este estudo oferece um passo nessa direção.

Agradecimentos

Agradecemos à Universidade do Estado do Amazonas (UEA), por meio do Programa de Produtividade Acadêmica 01.02.011304.026472/2023-87. Agradecemos também à

Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) pela concessão de bolsa de Iniciação Científica à primeira autora, no âmbito do Programa de Apoio à Iniciação Científica do Amazonas (PAIC-AM), Resolução n° 003/2025. Agradecemos especialmente aos estudantes participantes que contribuíram voluntariamente para este estudo.

Declaração sobre uso de Inteligência Artificial

Durante a preparação deste trabalho, a autora (pesquisadora) utilizou um modelo de linguagem de grande escala (ChatGPT, versão 5.2) apenas para fins limitados de suporte à escrita acadêmica. Especificamente, a ferramenta foi empregada para auxiliar na correção ortográfica e gramatical, e para validação da estrutura e clareza de alguns trechos do texto. A concepção do estudo, o desenho metodológico, a análise de dados, a interpretação dos resultados, a discussão das implicações e a redação principal do manuscrito foram realizados integralmente pela autora. A ferramenta de IA atuou estritamente como auxiliar na etapa final de revisão e polimento textual, sem contribuir para o conteúdo intelectual ou as conclusões da pesquisa. A autora revisou e editou criticamente todo o conteúdo gerado pela ferramenta e assume total responsabilidade pelo trabalho apresentado.

Referências

Alagarsamy, S., Sridhar, V., Krishnan, R., and Nandagopal, M. (2025). Enhancing large language models for text-to-testcase generation. *Information and Software Technology*, 180:107625.

Aranda, M., Oliveira, N., Soares, E., Ribeiro, M., Romão, D., Patriota, U., and Machado, I. (2024). A catalog of transformations to remove smells from natural language tests. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering (EASE 2024)*, pages 7–16.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Cohn, M. (2004). *User Stories Applied: For Agile Software Development*. Addison-Wesley Professional.

Cohn, M. (2024). User stories and user story examples. Mountain Goat Software.

Crompton, H. (2023). Artificial intelligence in higher education: The state of the field. *Computers and Education: Artificial Intelligence*, 4:100160.

Dutta, S. and Bhowmick, S. S. (2025). User stories: Does chatgpt do it better? In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025)*, pages 167–178.

Haldar, S., Pierce, M., and Capretz, L. F. (2025). Exploring the integration of generative AI tools in software testing education: A case study on ChatGPT and Copilot for preparatory testing artifacts in postgraduate learning. *IEEE Access*, 13:46070–46090.

Jalil, S., Rafi, S., LaToza, T. D., Moran, K., and Lam, W. (2023). ChatGPT and software testing education: Promises & perils. In *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pages 4130–4137. IEEE.

Kuhail, M. A., Farooq, S., Hammad, R., and Bahsoon, R. (2022). User story quality in practice: A case study. *Journal of Systems and Software*, 188:111269.

Li, Y., Wang, S., and Nguyen, T. N. (2025). Evaluating large language models for software testing. *Journal of Systems and Software*, 222:112345.

Lima, J. B. and Lima, M. S. (2025). Supplementary materials: Experimental evaluation of chatgpt for test case generation and quality analysis through test smells in software testing education. Figshare Repository.

Manzoni, F. S., Rodrigues, R., and Rocha, A. C. O. (2024). Exploring the use of chatgpt for the generation of user story based test cases: An experimental study. In *Proceedings of the XXXVIII Brazilian Symposium on Software Engineering (SBES 2024)*.

Mezzaro, S., Gambi, A., and Fraser, G. (2024). An empirical study on how large language models impact software testing learning. In *Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering (EASE 2024)*, pages 555–564. ACM.

Misu, M. R. H., Ahasan, K., Rahman, A., and Sakib, K. (2025). Test smell: A parasitic energy consumer in software testing. *Empirical Software Engineering*, 30(2):1–42.

Ouédraogo, W. C., Li, Y., Kaboré, K., Tang, X., Koyuncu, A., Klein, J., Lo, D., and Bissonandé, T. F. (2024). Test smells in LLM-generated unit tests. arXiv preprint arXiv:2410.10628.

Pitts, G., Marcus, V., and Motamedi, S. (2025). Student perspectives on the benefits and risks of AI in education. arXiv preprint arXiv:2502.01715.

Queiroz, F. K. and Lima, M. S. (2025). Uso do chatgpt na priorização de requisitos: Uma experiência educacional em engenharia de software. In *Anais do XIV Congresso Brasileiro de Informática na Educação (EduComp 2025)*.

Raharjana, I. K., Siahaan, D., and Fatichah, C. (2021). User stories and natural language processing: A systematic literature review. *IEEE Access*, 9:53811–53826.

Rodrigues, R., Manzoni, F. S., and Rocha, A. C. O. (2024). Exploring the use of large language models in requirements engineering education: An experience report with chatgpt 3.5. In *Anais do XXXVI Simpósio Brasileiro de Engenharia de Software (SBES 2024)*.

Santana Jr., E. G., Santos Junior, J. P., Almeida, E. P., Ahmed, I., Silveira Neto, P. A. M., and Almeida, E. S. (2025). Evaluating llms effectiveness in detecting and correcting test smells: An empirical study. In *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE 2025)*.