Evidências sobre o uso do ChatGPT no ensino de modelagem de software: um experimento controlado

Samir B. Murad[1], **Fabricio F. S. Lemos**[1], **Silvana M. Melo**[1], **Leo Natan Paschoal**[2], **Jorge M. Prates**[3]

[1]Universidade Federal da Grande Dourados (UFGD)
Dourados – MS – Brasil

[2]Pontifícia Universidade Católica do Paraná (PUCPR)
Curitiba – PR – Brasil

[3]Universidade Estadual de Mato Grosso do Sul (UEMS)
Dourados – MS – Brasil

{samirbmurad,fogacafabricioff}@gmail.com, silvanamelo@ufgd.edu.br

leo.paschoal@pucpr.br, jprates@uems.br

Abstract. *Chatbots LLMs have been explored across various domains and for different purposes, including as mechanisms for supporting education. When utilized as educational resources, it is essential to understand the effects of their use. This study describes a controlled experiment that analyzed the effects of using ChatGPT 3.5 in supporting software modeling education, specifically in the construction of UML diagrams. The experiment was designed to evaluate the effectiveness and efficiency of students in creating use case, class, and activity diagrams while also assessing the learning gains facilitated by using this resource. The results indicated that students who utilized ChatGPT demonstrated, on average, greater effectiveness and efficiency in producing the models. Furthermore, these students exhibited superior learning gains compared to those who engaged in modeling without the support of ChatGPT.*

Resumo. *Chatbots baseados em LLM têm sido explorados em diversos domínios e com diferentes propósitos, incluindo como mecanismos de apoio ao ensino. Quando utilizados como recursos educacionais, é essencial compreender os efeitos de seu uso. Este trabalho descreve um experimento controlado que analisou os efeitos da utilização do ChatGPT 3.5 no apoio ao ensino de modelagem de software, especificamente na construção de diagramas UML. O experimento foi elaborado com o objetivo de avaliar a eficácia e a eficiência dos estudantes na elaboração de diagramas de casos de uso, classes e atividades, verificando também o ganho de aprendizagem proporcionado pelo uso desse recurso. Os resultados indicaram que os estudantes que utilizaram o ChatGPT demonstraram em média uma maior eficácia e eficiência na produção dos modelos. Além disso, esses estudantes apresentaram um ganho de aprendizagem superior em comparação àqueles que realizaram a modelagem sem o suporte do ChatGPT.*

1. Introdução

Chatbots baseados em grandes modelos de linguagem (*Large Language Model - LLM*), como o *ChatGPT*[1] têm sido explorados em diversos contextos educacionais devido à capacidade de sintetizar grandes volumes de texto e gerar conteúdos de forma eficiente [Baidoo-Anu and Ansah 2023]. Revisões recentes têm indicado que o *ChatGPT*, por exemplo, tem sido utilizado em uma ampla gama de áreas, como saúde [Eysenbach et al. 2023], e também no ensino de tópicos relacionados à computação [Kosar et al. 2024].

Os estudos sobre o uso do *ChatGPT* como mecanismo de apoio ao ensino de computação focam na temática de programação [Kosar et al. 2024, Li et al. 2023, Khojah et al. 2024, Xue et al. 2024] envolvendo disciplinas como Algoritmos, Fundamentos de Programação e Programação Orientada a Objetos. No campo da Engenharia de Software, alguns esforços também têm sido realizados.

No ensino de Engenharia de Software, os estudos sobre o uso de grandes modelos de linguagem aplicados ao ensino investigam a capacidade desses modelos em fornecer resultados significativos quanto à eficácia dos estudantes [Choudhuri et al. 2024, Xue et al. 2024]. Além disso, existem pesquisas que realizam testes para avaliar a capacidade das ferramentas de produzir respostas concisas para questões didáticas [Rodriguez-Echeverría et al. 2024], bem como estudos que examinam o desempenho de sistemas baseados em LLM na oferta de respostas corretas e explicações precisas para perguntas comuns relacionadas a conteúdos de Engenharia de Software [Jalil et al. 2023].

Alguns estudos também têm avaliado a eficácia do *ChatGPT* como suporte à elaboração de modelos em linguagem UML (*Unified Modeling Language*) [Russo 2024, Schäfer et al. 2024, Abukhalaf et al. 2023], analisando a corretude, tanto sintática quanto semântica dos modelos *UML* gerados. No entanto, essa capacidade ainda não foi explorada no contexto educacional, especialmente no que diz respeito à capacidade desse sistema orientar alunos a elaborar diagramas em *UML* com eficácia e eficiência.

Este trabalho tem como objetivo avaliar a utilização do *ChatGPT* no contexto de ensino de modelagem de software, com ênfase em projetos de software que utilizam modelagem orientada a objetos em linguagem *UML*. A pesquisa foi conduzida por meio de um experimento controlado [Wohlin et al. 2012], com o intuito de avaliar a eficácia, a eficiência e o ganho de aprendizagem ao utilizar o *ChatGPT* como auxílio na produção de modelos *UML* em atividades educacionais.

O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados a esta pesquisa; a Seção 3 descreve o planejamento e a condução do estudo experimental; a Seção 4 relata os resultados do experimento; e, por fim, a Seção 5 traz as conclusões e sugestões para trabalhos futuros.

2. Trabalhos relacionados

A fim de identificar estudos que abordassem avaliações empíricas sobre o uso de LLM na área de Engenharia de Software, foi realizada uma revisão da literatura. Os estudos identificados abrangem diferentes áreas da Engenharia de Software, como: (i) ensino e treinamento de Engenharia de Software; (ii) engenharia de requisitos; (iii) teste de software; (iv) análise e modelagem de software; (v) estudos secundários; (vi) desafios e perspectivas da área.

[1]Mais informações disponíveis em: https://chatgpt.com/

No contexto de ensino de Engenharia de Software, alguns estudos têm avaliado o auxílio do *ChatGPT* em cursos de computação, discutindo como os LLMs podem ser incluídos no ensino de Engenharia de Software auxiliando os estudantes na elaboração, depuração e tradução de códigos entre linguagens, análise de desafios éticos e sua integração na educação, bem como auxílio à resolução de atividades avaliativas [Golgiyaz 2023, Rodriguez-Echeverría et al. 2024].

No que tange ao ensino de engenharia de requisitos, estudos têm avaliado o uso do *ChatGPT* principalmente na criação de melhores *prompts* para a criação e classificação de requisitos (funcionais, não-funcionais e de negócio) de software mais consistentes. Os resultados indicam que o uso de LLM pode ser útil nesse tópico para diminuir a carga de tarefas, fornecendo uma base para um projeto e listando requisitos que podem ser aplicados e aperfeiçoados no projeto [Ronanki et al. 2023], [Sami et al. 2024], [Ronanki et al. 2024].

Na indústria, para o treinamento de profissionais em Engenharia de Software, estudos empíricos têm proposto o uso do *ChatGPT* em atividades diárias de programação, testes, gerenciamento de projeto, engenharia de requisitos e prototipação [Khojah et al. 2024]. Os resultados indicam que o *ChatGPT* pode ser útil para aprender novos conceitos e tomar decisões de projeto, além de reduzir tarefas repetitivas que não sejam complexas.

Na temática de modelagem e análise de software, os estudos encontrados abordam o uso de LLM como auxílio à construção de modelos *UML*. Os estudos de [Abukhalaf et al. 2023] e [Chaaben et al. 2023] abordam o uso de *prompts* para autocompletar diagramas de atividades, diagramas de classes e outros, demonstrando a eficiência do *ChatGPT* para essas tarefas. O estudo de [Cámara et al. 2023] propõe além dos conceitos de modelagem, o uso da linguagem de marcação *PlantUML* em conjunto com o *ChatGPT* para a geração dos modelos, auxiliando desse modo a visualização dos diagramas que compõem os projetos de Engenharia de Software.

Embora existam trabalhos preliminares sobre o ensino de Engenharia de Software apoiado pelo *ChatGPT* e existam estudos que avaliem a capacidade do *ChatGPT* auxiliar na tarefa de elaboração de diagramas *UML* para representar a estrutura e o comportamento de sistemas de software, em relação à corretude e qualidade desses modelos [Choudhuri et al. 2024, Russo 2024, Schäfer et al. 2024, Abukhalaf et al. 2023], não há estudos que avaliem o uso desse recurso no processo de ensino-aprendizagem. Portanto, este trabalho visa contribuir com o entendimento a cerca da eficácia, eficiência e ganho de aprendizagem proporcionado pelo *ChatGPT*, quando utilizado no contexto do ensino de atividades de modelagem de diagramas *UML*.

3. Materiais e métodos

Esta seção apresenta o contexto de condução da pesquisa. O estudo adota a metodologia proposta por [Wohlin et al. 2012] para a realização de estudos experimentais, e as etapas desse processo são detalhadas nas próximas seções.

3.1. Definição do escopo

Conforme a abordagem GQM (*Goal, Question, Metrics*) [Basili et al. 1994], o escopo do experimento relatado neste trabalho pode ser descrita da seguinte forma: **analisar** o *ChatGPT* como mecanismo de apoio à realização de atividades educacionais de modelagem de software, **com o propósito de** verificar, **com respeito à** eficácia, eficiência e

ganho de aprendiagem, **do ponto de vista de** pesquisadores, **no contexto de** alunos de graduação que estavam realizando uma atividade educacional relacionada a elaboração de diagrama de classes, casos de uso e diagrama de atividades em *UML*.

Para compreender os efeitos de uso do *ChatGPT* durante a realização de atividades educacionais sobre modelagem de software, foram estipuladas três questões de pesquisa e um par de hipóteses (HN - hipótese nula e HA- hipótese alternativa) para cada questão de pesquisa (QP), conforme apresentado na Tabela 1.

Tabela 1. Questões de pesquisa e hipóteses do experimento

QP	Hipótese nula	Hipótese alternativa
QP 1: O uso do *ChatGPT* melhora a eficácia dos estudantes na modelagem *UML*?	HN0: Não há diferença significativa na eficácia entre os alunos que modelam diagramas *UML* com apoio do *ChatGPT* e aqueles que modelam sem esse suporte.	HA0: Há diferença significativa na eficácia entre estudantes que modelam diagramas *UML* com o suporte do *ChatGPT* e aqueles que modelam sem esse suporte.
QP 2: O uso do *ChatGPT* melhora a eficiência dos estudantes na modelagem de diagramas *UML*?	HN1: Não há diferença significativa na eficiência entre estudantes que modelam diagramas *UML* com o suporte do *ChatGPT* e aqueles que modelam sem esse suporte.	HA1: Há diferença significativa na eficiência entre estudantes que modelam diagramas *UML* com o suporte do *ChatGPT* e aqueles que modelam sem esse suporte.
QP 3: O uso do *ChatGPT* proporciona um maior ganho de aprendizagem na construção de diagramas *UML*?	HN2: Não há diferença significativa no ganho de aprendizagem entre estudantes que utilizam o *ChatGPT* para modelagem de diagramas *UML* e aqueles que não utilizam.	HA2: Há diferença significativa no ganho de aprendizagem entre estudantes que utilizam o *ChatGPT* para modelagem de diagramas *UML* e aqueles que não utilizam.

3.2. Seleção das variáveis

Para identificar as variáveis do estudo, o framework estabelecido no trabalho de [Paschoal 2024] foi utilizado. Nesse sentido, o experimento contempla variáveis independentes e dependentes. A variável independente considerada no experimento é o mecanismo de apoio ao ensino. Entende-se por mecanismos de apoio ao ensino recursos, estratégias e abordagens utilizadas para facilitar a aprendizagem de determinado conteúdo. Assim, neste experimento, essa variável possui dois tratamentos, destacados a seguir:

- **Tratamento A**: o *ChatGPT* está disponível para que os alunos da disciplina o utilizem na produção de diagramas em *UML*.
- **Tratamento B**: o *ChatGPT* não está disponível, de modo que os alunos desenvolvam os diagramas em *UML* sem qualquer apoio adicional.

O objetivo do experimento é investigar os efeitos da utilização do *ChatGPT* em três variáveis dependentes:

- **Eficácia**: grau em que os alunos conseguem representar corretamente os elementos necessários no diagrama em *UML*, conforme a descrição de uma atividade proposta e o tipo de diagrama. Nesse sentido, mede-se a correspondência entre os elementos modelados e os requisitos da atividade.
- **Eficiência**: remete ao tempo (em minutos) que os participantes levam para realizar a atividade proposta.
- **Ganho de aprendizagem**: refere-se ao progresso ou avanço de conhecimento ou habilidades de um estudantes após as instruções e execução das atividades propostas.

3.3. Amostragem e design do experimento

Para obter evidências sobre o uso do *ChatGPT* no ensino de modelagem, o estudo requer a seleção de estudantes e a elaboração de diagramas *UML* utilizando os mecanismos de apoio estabelecidos em cada tratamento. Para isso, foi necessário definir a amostragem e uma estratégia de alocação dos participantes. Por conveniência, foram selecionados alunos de graduação da Universidade Federal da Grande Dourados (UFGD), matriculados nos cursos de Engenharia de Computação e Sistemas de Informação e na disciplina de Engenharia de Software. Optou-se por um design experimental independente, no qual os estudantes são selecionados para participar do estudo e distribuídos aleatoriamente em um dos tratamentos, utilizando apenas aquele ao qual foram designados.

3.4. Instrumentos para a condução

Considerando a variável independente em destaque neste experimento, a natureza da atividade de modelagem de software e as variáveis dependentes, foi necessário preparar instrumentos para viabilizar a execução do experimento e a coleta de dados.

O *ChatGPT*, objeto de estudo, não é capaz de gerar modelos *UML* diretamente [Cámara et al. 2023]. No entanto, ele pode elaborar códigos de marcação, e algumas ferramentas de modelagem de software possibilitam a criação de diagramas *UML* a partir desses códigos, como a *PlantUML*. Diante disso, para permitir o uso do *ChatGPT* na modelagem de software, os alunos submetidos ao tratamento A tiveram acesso à ferramenta, podendo utilizá-la para realizar atividades educacionais de modelagem em *UML* por meio da *PlantUML*.

Para garantir que os alunos atribuídos ao tratamento A conseguissem modelar utilizando a *PlantUML*, foi necessário preparar e oferecer um treinamento preliminar. Esse treinamento foi semelhante a uma aula tradicional de modelagem, mas, em vez de aprenderem a modelar com ferramentas visuais, os alunos aprenderam a utilizar o *PlantUML*.

Os alunos submetidos ao tratamento B não podem usar o *ChatGPT* para esclarecer dúvidas sobre a elaboração de modelos *UML* ou solicitar sugestões de modelagem. Dessa forma, ficou decidido que esses alunos aprenderiam a modelar utilizando a ferramenta Visual Paradigm, recebendo um treinamento específico sobre seu uso.

Com os materiais definidos para a execução dos tratamentos, foi necessário estabelecer atividades que exigissem dos alunos a modelagem de diagramas *UML*. No estudo foram definidas três atividades: a primeira envolvia a modelagem de um diagrama de casos de uso (Atividade 1); a segunda, a modelagem de um diagrama de classes (Atividade 2); e, por fim, a terceira exigia a modelagem de um diagrama de atividades (Atividade 3). Essas atividades foram planejadas para permitir uma análise específica sobre os efeitos do uso de diferentes recursos na realização de tarefas educacionais em diagramas *UML* específicos.

A decisão de limitar as atividades a três diagramas foi motivada pela necessidade de evitar sobrecarga nos alunos, que poderia ocorrer caso todos os 14 tipos de diagramas *UML* fossem exigidos. Vale destacar que as atividades foram utilizadas como base para avaliar a eficácia dos alunos na modelagem de diagramas *UML*.

A Figura 1 ilustra as etapas de condução de cada atividade, que consiste em fornecer a especificação (definição) de um sistema de software, em seguida o aluno deve modelar o diagrama solicitado e, por fim, após a entrega da atividade, esta é corrigida pelos pesquisadores do estudo.

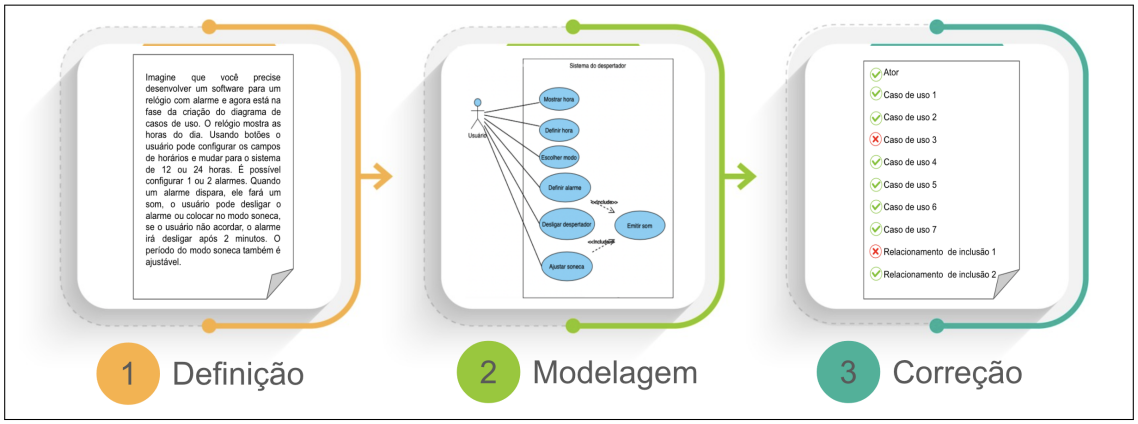

Figura 1. Etapas da execução das atividades.

3.5. Execução do experimento

O experimento foi conduzido entre abril e outubro de 2024, com alunos matriculados na disciplina de Engenharia de Software I, dos cursos de Engenharia de Computação e Sistemas de Informação, nos laboratórios de informática da Faculdade de Ciências Exatas e Tecnologia (FACET) da UFGD. A execução ocorreu em três etapas, cada uma realizada em um momento distinto da disciplina. Como consequência, o número de participantes variou ao longo do experimento.

A primeira etapa do experimento ocorreu com uma amostra de 13 alunos do curso de Sistemas de Informação e envolveu a realização da Atividade 1, na qual os alunos deveriam elaborar individualmente diagramas de casos de uso. Esses diagramas representam os usuários ou atores e suas interações com o sistema, incluindo fluxos de ações principais e alternativos.

Antes da realização dessa etapa, todos os alunos da disciplina haviam participado de uma aula teórica no modelo tradicional de ensino, com explicação e esclarecimento de dúvidas conduzidos pela professora responsável pela disciplina em sala de aula.

Os alunos foram alocados aleatoriamente nos tratamentos do experimento, tomando como base o número de alunos que concordaram com o TCLE. Para formar os grupos, considerou-se a quantidade de alunos presentes na aula no dia da realização do experimento.

A segunda etapa ocorreu com uma amostra de 23 alunos, também do curso de Sistemas de Informação, e envolveu a realização da atividade de modelagem de diagramas de classe, que representam a estrutura de classes, atributos, métodos e suas relações.

De maneira semelhante à primeira etapa, após uma aula teórica sobre o tema, os alunos foram alocados aos tratamentos para realizar a modelagem do diagrama de classes individualmente. Os estudantes atribuídos ao tratamento A utilizaram o *ChatGPT* em conjunto com o *PlantUML* para auxiliar na criação dos diagramas, enquanto os alunos alocados ao tratamento B utilizaram exclusivamente a ferramenta CASE convencional.

A terceira etapa ocorreu com uma amostra de 22 alunos do curso de Engenharia de Computação e abordou a elaboração de diagramas de atividades, que permitem simular o funcionamento de um sistema utilizando a notação *UML*.

De maneira semelhante às etapas anteriores, após os alunos participarem de uma aula teórica sobre esse tipo de diagrama, eles foram alocados aos tratamentos para realizar as atividades de modelagem de forma individual, utilizando as mesmas ferramentas

disponibilizadas anteriormente, tendo como base o tratamento para o qual foi alocado.

O diferencial desta etapa em relação às demais foi a aplicação de testes de conhecimento com os participantes da pesquisa, com o objetivo de analisar o conhecimento adquirido sobre o assunto abordado. Dessa forma, o ganho de aprendizagem foi mensurado exclusivamente com base no aprendizado relacionado à modelagem de diagramas de atividades.

4. Resultados e Discussões

Esta seção apresenta os principais resultados e discussões a fim responder as questões de pesquisa acerca da adoção do *ChatGPT* no ensino de modelagem de software. Cada uma da variáveis avaliadas é descrita em detalhes nas próximas seções.

Vale salientar que, durante a apresentação dos resultados, os dados dos participantes vinculados ao tratamento A serão considerados como pertencentes ao grupo experimental, enquanto os dados dos participantes vinculados ao tratamento B serão classificados como grupo controle. Essa distinção entre grupo experimental e grupo controle é usada para facilitar a comparação dos efeitos dos tratamentos, possibilitando identificar eventuais diferenças nos resultados decorrentes da aplicação dos diferentes tratamentos.

4.1. Eficácia

A análise da variabilidade dos dados relacionados à eficácia (taxa de acertos - variando de 0 a 10 pontos) nas atividades de modelagem, entre os grupos experimentais e controle, apresentada no gráfico boxplot da Figura 2, mostra que em média os alunos que utilizaram o *ChatGPT* (grupo experimental) alcançaram uma nota mais alta que o grupo que não utilizou (grupo controle). Porém a diferença é significativa apenas para as atividades 1 e 2, conforme provado pelos testes estatísticos de hipótese.

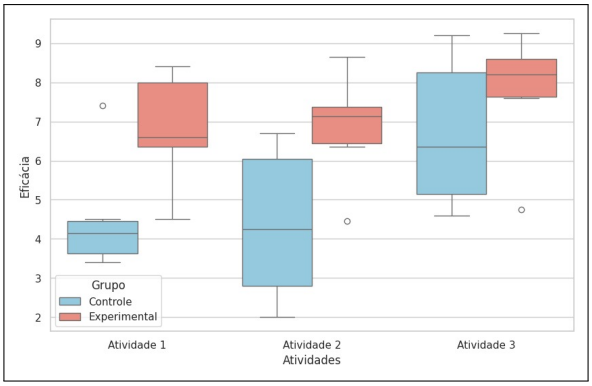

Figura 2. Distribuição da eficácia entre os grupos

A análise descritiva dos dados relacionados à eficácia para cada uma das atividades é apresentada na Tabela 2. Conforme indicado nessa tabela, a verificação da normalidade dos dados foi realizada por meio do teste de Shapiro-Wilk, que resultou em valores de p iguais a 0,659, 0,260 e 0,907 para as atividades 1, 2 e 3, respectivamente. Esses resultados indicam que as amostras seguem uma distribuição normal, o que justifica o uso do teste t de Student para a análise das hipóteses.

A análise da hipótese nula (HN0), definida na Seção 3.1, afirma que não há diferença significativa de eficácia entre os alunos que utilizam o *ChatGPT* como auxílio na realização das atividades e aqueles que não utilizam. No entanto, os resultados dos

Tabela 2. Análise estatística da Eficácia

Atv.	Grupo	N	Média	Mediana	Desvio-padrão	Erro-padrão	Shapiro-Wilk (p-valor)	Teste t (p-valor)
1	Controle	6	4,52	4,15	1,48	0,603	0,659	0,012
	Experimental	7	6,89	6,60	1,37	0,518		
2	Controle	13	4,50	4,25	1,72	0,477	0,260	0,001
	Experimental	10	6,88	7,13	1,08	0,341		
3	Controle	9	6,70	6,35	1,73	0,577	0,907	0,145
	Experimental	8	7,88	8,20	1,39	0,493		

testes estatísticos indicam que a hipótese nula pode ser rejeitada para as atividades 1 e 2, com valores de p iguais a 0,012 e 0,001, respectivamente. Por outro lado, para a atividade 3, não foi possível rejeitar a hipótese nula, pois o valor de p foi igual a 0,145. Desse modo, há diferenças significativas na eficácia dos alunos nas atividades de modelagem de diagramas de casos de uso e diagramas de classes, mas não há diferenças significativas para atividades relacionadas à modelagem de diagramas de atividades, considerando as amostras analisadas.

4.2. Eficiência

A análise da eficiência revelou diferenças significativas no tempo necessário para a conclusão das atividades entre os alunos que utilizaram o *ChatGPT* (grupo experimental) e aqueles que não utilizaram (grupo controle). A Figura 3 apresenta um gráfico de boxplot que ilustra a distribuição do tempo médio gasto por cada grupo na resolução das atividades. Observa-se que o grupo experimental, que contou com o auxílio do *ChatGPT* para a realização das atividades, gastou em média 57 minutos a menos na primeira atividade e 50 minutos a menos na segunda atividade. Essa diferença é especialmente relevante, considerando que a duração total máxima das atividades era de 120 minutos.

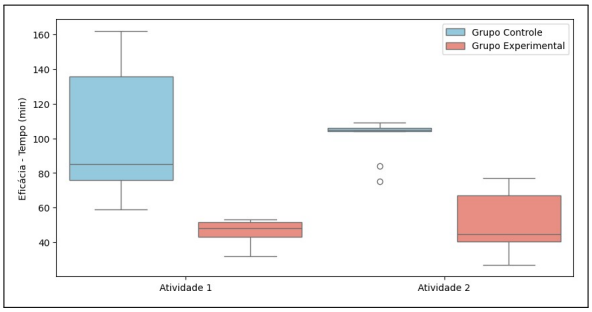

Figura 3. Distribuição da eficiência dos grupos

A Tabela 3 apresenta os dados estatísticos da análise de eficiência entre os grupos. O teste de Shapiro-Wilk indicou que a distribuição dos dados é normal para a atividade 1, com valores de p iguais a 0,1528 e 0,0865, respectivamente. Para a atividade 2, os valores de p iguais a 0,0001 e 0,2883 indicam que a distribuição não é normal.

Vale salientar que, na atividade 3, houve um problema na coleta dos dados de tempo de início da modelagem do diagrama de atividades, o que impossibilitou o cálculo da eficiência.

Em razão das distribuição dos dados, o teste de hipóteses foi conduzido utilizando o teste t de Student para a atividade 1 e o teste de Mann-Whitney para a atividade 2. Na Tabela 3, o indicador 'NA' (não aplicável) é utilizado para os conjuntos de dados aos quais o teste estatístico não se aplica.

Tabela 3. Análise estatística da Eficiência

Ativ.	Grupo	N	Média (minutos)	Mediana (minutos)	Desvio-padrão	Shapiro-Wilk (p-valor)	Teste t (p-valor)	Mann-Whitney (p-valor)
1	Controle	6	102,67	85,00	42,99	0,1528	0,0226	NA
	Experimental	7	46,00	48,00	1,37	0,0865		
2	Controle	13	101,69	105,00	10,13	0,0001	NA	0,0001
	Experimental	10	51,30	44,50	16,91	0,2883		

A hipótese nula (HN1) sobre a eficiência, que afirma que não há diferença significativa no tempo de realização das atividades entre os alunos que utilizaram o *ChatGPT* e os que não utilizaram, pode ser rejeitada, uma vez que os resultados dos testes indicaram um valor de *p* igual a 0,0226 para o Teste *t* e um valor de *p* de 0,0001 para o Teste de Mann-Whitney, respectivamente. Dessa forma, observa-se que os alunos que utilizaram o *ChatGPT* para modelar os diagramas de casos de uso e diagramas de classes levaram menos tempo do que os alunos que não tiveram acesso a essa ferramenta.

4.3. Ganho de Aprendizagem

O ganho de aprendizagem obtido pelos estudantes a partir da realização das atividades educacionais foi calculado levando em consideração o conhecimento inicial de cada aluno. A premissa subjacente é que um aluno com um desempenho mais alto no pré-teste já possui um nível de conhecimento mais avançado, e, portanto, tem menos espaço para melhorar em comparação a um aluno com um desempenho mais baixo. Dessa forma, a equação usada para calcular o ganho de aprendizagem, apresentada a seguir, leva em conta a diferença entre o desempenho no pós-teste e no pré-teste, ajustada pelo desempenho inicial do aluno, o que permite uma avaliação mais justa e equitativa do progresso [Paschoal 2024].

$$Ganho\ de\ aprendizagem = \frac{\text{Pós-teste} - \text{Pré-teste}}{1 - \text{Pré-teste}}$$

No gráfico de boxplot apresentado na Figura 4, é possível observar o aumento no ganho de aprendizagem para os alunos do grupo experimental. A análise estatística dos resultados confirma a diferença significativa no ganho de aprendizagem entre os grupos controle e experimental, como pode ser observado na Tabela 4. Os alunos do grupo experimental apresentaram um ganho máximo de 0,695, enquanto no grupo controle esse ganho foi de -0,318. Esse valor negativo para o grupo controle sugere que, em média, os alunos desse grupo não tiveram uma melhoria no desempenho, e alguns até tiveram um desempenho inferior no pós-teste, o que reforça a diferença significativa entre os dois grupos.

Tabela 4. Análise estatística do Ganho de Aprendizagem

Grupo	N	Média	Mediana	Desvio-padrão	Mínimo	Máximo	Shapiro-Wilk (p-valor)	Teste t (p-valor)
Controle	8	-0,508	-0,497	0,147	-0,775	-0,318	0,729	$<.001$
Experimental	8	0,274	0,313	0,333	-0,284	0,695	0,487	

Por meio da estatística descritiva e do teste de Shapiro-Wilk, foi possível comprovar a normalidade da amostra, o que permite a aplicação do Teste *t*. O resultado foi um p-valor $< 0{,}001$, o que indica que é possível rejeitar a hipótese nula (HN2), que afirma que não há diferenças significativas no ganho de aprendizagem entre os alunos que utilizam o *ChatGPT* na resolução das atividades e os que não utilizam.

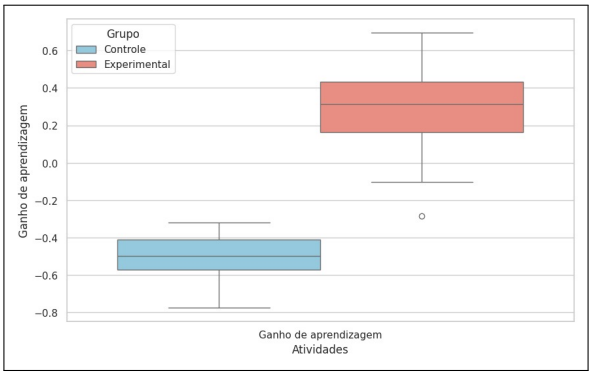

Figura 4. Distribuição do ganho de aprendizagem entre os grupos

Esses resultados sugerem que a utilização do *ChatGPT* pode, além de auxiliar na rapidez de conclusão e na corretude das atividades, proporcionar um maior nível de retenção de conhecimento pelos alunos, aumentando, portanto, o ganho de aprendizagem durante a execução das atividades de modelagem.

5. Conclusões

Este estudo apresentou uma análise da eficácia, eficiência e ganho de conhecimento decorrente do uso do *ChatGPT* no ensino de modelagem de software com diagramas *UML*, no contexto da disciplina de Engenharia de Software em cursos de graduação em Computação. Os resultados obtidos indicam que o uso do *ChatGPT* durante a realização das atividades de modelagem pode trazer benefícios aos estudantes, tanto no aumento da taxa de acerto dos elementos que compõem o diagrama quanto na redução do tempo de execução das atividades e no ganho de aprendizagem.

Apesar dos resultados, em sua maioria, serem positivos, há necessidade de conduzir estudos adicionais sobre o uso do *ChatGPT* como apoio ao ensino de Engenharia de Software, a fim de validar as conclusões obtidas e aprofundar a compreensão sobre seu impacto na educação, especialmente em comparação com o método tradicional de ensino.

Uma limitação deste estudo é o fato de considerar apenas um tópico dentro da disciplina de Engenharia de Software, tendo sido conduzido em duas turmas de dois cursos de Computação de uma mesma universidade, com um número restrito de alunos. Para a generalização dos resultados, seria importante replicar o estudo em diferentes cenários, com amostras de tamanho variado, possibilitando a comparação dos resultados obtidos e gerando novas evidências.

Além disso, como oportunidade para o desenvolvimento de pesquisas futuras, pode-se replicar o estudo em outras instituições de ensino e considerando outros diagramas *UML*, como diagramas de sequência e fluxo de dados. Também é possível investigar os efeitos do uso do *ChatGPT* na percepção dos alunos durante a modelagem de software com *UML*, incluindo aspectos como a utilidade percebida da ferramenta, a aceitação da tecnologia e os impactos na carga cognitiva dos estudantes.

Agradecimentos

Os autores gostariam de agradecer ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná (PUCPR) e a PROPP/UFGD - SIGProj nº 322855.1174.8276.11032019.

Referências

Abukhalaf, S., Hamdaqa, M., and Khomh, F. (2023). On codex prompt engineering for ocl generation: An empirical study. In *2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR)*, pages 148–157.

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. *Journal of AI*, 7(1):52–62.

Basili, V. R., Caldiera, G., and Rombach, D. H. (1994). *The Goal Question Metric Approach*, volume I. John Wiley & Sons.

Cámara, J., Troya, J., Burgueño, L., and Vallecillo, A. (2023). On the assessment of generative ai in modeling tasks: an experience report with chatgpt and uml. *Softw. Syst. Model.*, 22(3):781–793.

Chaaben, M. B., Burgueño, L., and Sahraoui, H. (2023). Towards using few-shot prompt learning for automating model completion. In *Proceedings of the 45th International Conference on Software Engineering: New Ideas and Emerging Results*, ICSE-NIER '23, page 7–12. IEEE Press.

Choudhuri, R., Liu, D., Steinmacher, I., Gerosa, M., and Sarma, A. (2024). How far are we? the triumphs and trials of generative ai in learning software engineering. In *2024 IEEE/ACM 46th International Conference on Software Engineering (ICSE)*, ICSE '24, pages 2270–2282, Los Alamitos, CA, USA. IEEE Computer Society.

Eysenbach, G. et al. (2023). The role of chatgpt, generative language models, and artificial intelligence in medical education: a conversation with chatgpt and a call for papers. *JMIR medical education*, 9(1):e46885.

Golgiyaz, S. (2023). Chatgpt in computer software education. *ICHEAS, 4th International Conference On Health, Engineering And Applied Sciences.*, pages 115–126.

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.

Khojah, R., Mohamad, M., Leitner, P., and de Oliveira Neto, F. G. (2024). Beyond code generation: An observational study of chatgpt usage in software engineering practice. *Proc. ACM Softw. Eng.*, 1(FSE).

Kosar, T., Ostojić, D., Liu, Y. D., and Mernik, M. (2024). Computer science education in chatgpt era: Experiences from an experiment in a programming course for novice programmers. *Mathematics*, 12(5).

Li, Y., Xu, J., Zhu, Y., Liu, H., and Liu, P. (2023). The impact of chatgpt on software engineering education: A quick peek. In *2023 10th International Conference on Dependable Systems and Their Applications (DSA)*, pages 595–596.

Paschoal, L. N. (2024). *Um framework para o planejamento de experimentos controlados na pesquisa de chatbots educacionais*. Tese de doutorado, Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos.

Rodriguez-Echeverría, R., Gutiérrez, J. D., Conejero, J. M., and Prieto, A. E. (2024). Analysis of chatgpt performance in computer engineering exams. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 19:71–80.

Ronanki, K., Berger, C., and Horkoff, J. (2023). Investigating chatgpt's potential to assist in requirements elicitation processes. In *2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 354–361.

Ronanki, K., Cabrero-Daniel, B., Horkoff, J., and Berger, C. (2024). *Requirements Engineering Using Generative AI: Prompts and Prompting Patterns*, pages 109–127. Springer Nature Switzerland, Cham.

Russo, D. (2024). Navigating the complexity of generative ai adoption in software engineering. *ACM Trans. Softw. Eng. Methodol.*, 33(5).

Sami, M. A., Rasheed, Z., Waseem, M., Zhang, Z., Herda, T., and Abrahamsson, P. (2024). Prioritizing software requirements using large language models. *CoRR*, abs/2405.01564.

Schäfer, M., Nadi, S., Eghbali, A., and Tip, F. (2024). An empirical evaluation of using large language models for automated unit test generation. *IEEE Transactions on Software Engineering*, 50(1):85–105.

Wohlin, C., Runeson, P., Hst, M., Ohlsson, M. C., Regnell, B., and Wessln, A. (2012). *Experimentation in Software Engineering*. Springer Publishing Company, Incorporated.

Xue, Y., Chen, H., Bai, G. R., Tairas, R., and Huang, Y. (2024). Does chatgpt help with introductory programming?an experiment of students using chatgpt in cs1. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training*, ICSE-SEET '24, page 331–341, New York, NY, USA. Association for Computing Machinery.