Avaliação de Conversas Educacionais Sintéticas Geradas por LLMs no Ensino de Programação

Flávio Izo[1], João Pedro Chaves Cruz[1], Lorenzo Rainha Gomes[1], Lucas Guimarães Bosio Altoé[1], Maria Eduarda Agum Mendonça Chagas[1]

[1]Instituto Federal do Espírito Santo (IFES)
Cachoeiro de Itapemirim – Espírito Santo – Brasil

fizo@ifes.edu.br, joaopedroccruz.dev@gmail.com, lorenzorgomes@gmail.com

lucas.galc@gmail.com, mariaeduardaagum@gmail.com

Abstract. *This work proposes and evaluates* SCEPA*, a multidimensional metric and framework for analyzing synthetic educational dialogues generated by Large Language Models (LLMs) in programming education. Four LLMs were compared across ten scenarios and evaluated by six raters using Likert-scale criteria. The analysis included inter-rater reliability and descriptive statistics. Results indicate differences in pedagogical performance among the models, with Gemini 2.5 Pro achieving the best score (4.714), followed by DeepSeek R1 (4.631) and Claude Sonnet 4 (4.552), while ChatGPT-5 obtained a lower score (3.791). These findings demonstrate the usefulness of the* SCEPA *index for evaluating educational dialogues generated by LLMs.*

Resumo. *Este trabalho propõe e avalia o* SCEPA*, uma métrica e estrutura multidimensional para análise de diálogos educacionais sintéticos gerados por Modelos de Linguagem de Grande Escala (LLMs) no ensino de programação. Quatro LLMs foram comparados em dez cenários e analisados por seis avaliadores utilizando critérios em escala Likert. A análise incluiu confiabilidade interavaliadores e estatísticas descritivas. Os resultados indicam diferenças no desempenho pedagógico entre os modelos, com Gemini 2.5 Pro com o melhor desempenho (4,714), seguido por DeepSeek R1 (4,631) e Claude Sonnet 4 (4,552), enquanto o ChatGPT-5 obteve pontuação inferior (3,791). Esses resultados demonstram a utilidade do índice* SCEPA *para avaliar diálogos educacionais gerados por LLMs.*

1. Introdução

A interação dialógica entre professor e estudante é reconhecida como um elemento central no processo de aprendizagem, favorecendo a construção de conhecimento e o engajamento discente. Pesquisas demonstram que abordagens lúdicas, incluindo o uso de histórias e conversas contextualizadas, podem facilitar a compreensão de conceitos abstratos e estimular o engajamento estudantil [M. Valença and Balthazar Tostes 2019, Bortolazzo 2024]. Entretanto, a aplicação de conversas em contextos educacionais pode enfrentar alguns desafios. Um obstáculo central reside na escassez de dados reais específicos e de qualidade que documentem interações autênticas professor-aluno. Nesse cenário, dados sintéticos emergem como alternativa para a pesquisa

empírica, permitindo investigações sobre a qualidade conversacional, eficácia pedagógica e aceitação estudantil.

Nesse contexto, os Modelos de Linguagem de Larga Escala (em inglês *Large Language Model* - LLMs) surgem como ferramentas promissoras para apoiar o processos educacionais, especialmente no ensino de programação, onde os estudantes frequentemente enfrentam dificuldades conceituais que demandam atenção individualizada. A capacidade desses modelos de gerar conversas naturais e contextualizadas abre possibilidades para simular interações pedagógicas entre professores e alunos, funcionando como material educacional complementar.

No contexto específico do ensino de programação, conversas sintéticas apresentam benefícios potenciais relevantes: estudantes com perfil mais tímido podem identificar suas próprias dúvidas refletidas nos diálogos apresentados, sentindo-se encorajados a questionar o professor; as dúvidas sintéticas podem parecer com questionamentos reais de outros estudantes, validando suas dificuldades; e os próprios alunos podem exercitar o papel de tutores ao responder às questões do personagem estudante sintético, consolidando seu conhecimento através do ensino.

Esta pesquisa propõe e valida o índice *Synthetic Conversational Educational Performance Assessment* (*SCEPA*, em português Avaliação do Desempenho Educacional em Conversas Sintéticas), um instrumento multidimensional composto por 8 critérios avaliativos: Precisão Técnica, Adequação e Metodologia Pedagógica, Realismo, Empatia, Consistência da Persona, Relevância, Aplicabilidade e Ortografia. Foi desenvolvido um *framework* flexível que permite a configuração contextual dos pesos de cada critério conforme necessidades pedagógicas específicas, oferecendo autonomia aos educadores na definição de prioridades avaliativas dos critérios das conversas sintéticas. Foram feitos experimentos empíricos envolvendo quatro LLMs muito utilizados (ChatGPT 5[1], Claude 4 Sonnet[2], Gemini 2.5 Pro[3] e DeepSeek R1[4]), 10 cenários educacionais de programação e 240 avaliações realizadas por seis avaliadores.

Este trabalho aborda o tripé da área de computação: Pessoas (educadores e estudantes que interagem com tecnologias educacionais), Processos (metodologias de avaliação de qualidade conversacional e procedimentos de integração de Inteligência Artificial (IA) em práticas pedagógicas) e Tecnologias (LLMs como ferramentas de apoio educacional). Esta pesquisa alinha-se aos Grandes Desafios da Computação no Brasil (2025-2035) [Sociedade Brasileira de Computação 2025], especialmente no que se refere ao uso responsável e eficaz da IA em contextos educacionais. Ao investigar a qualidade de conversas sintéticas geradas por LLMs, a pesquisa contribui para o desenvolvimento de abordagens que integrem IA ao processo de aprendizagem, considerando aspectos técnicos, pedagógicos e de interação humano-computador.

O estudo fundamenta-se teoricamente na perspectiva sociointeracionista de Vygotsky [Vygotsky 1978], especialmente no conceito de Zona de Desenvolvimento Proximal (ZDP), reconhecendo que os LLMs podem atuar como mediadores tecnológicos

[1]https://chatgpt.com/
[2]https://claude.ai/
[3]https://gemini.google.com/
[4]https://deepseek-portugues.chat/

no processo de aprendizagem quando adequadamente calibrados para contextos educacionais e assim atingir o Nível de Desenvolvimento Potencial (NDP). Nesse sentido, o papel do adulto ou colega mais experiente seria exercido pelas LLMs.

Trabalhos utilizando a IA no contexto educacional fornecem suporte ao aprendizado, unindo acessibilidade e personalização das tarefas, e consequentemente contribuindo para o avanço do ensino assistido por IA [Siqueira et al. 2025]. No entanto, apesar do crescente uso de LLMs na educação, ainda há pouca investigação sobre como avaliar sistematicamente a qualidade pedagógica de diálogos educacionais sintéticos gerados por esses modelos.

As contribuições deste trabalho incluem: (i) a proposição e validação do índice *SCEPA* para avaliação de conversas sintéticas educacionais; (ii) o desenvolvimento de um *framework* configurável para aplicação contextualizada do índice; (iii) análise comparativa sistemática de quatro LLMs no contexto educacional de programação; e (iv) disponibilização de *dataset* de conversas sintéticas para futuras pesquisas.

Visando a reprodutibilidade deste estudo, todos os materiais da pesquisa estão disponíveis publicamente[5], incluindo os questionários aplicados e as conversas geradas pelas *LLMs* e avaliadas seguindo os critérios do índice *SCEPA*.

2. Fundamentação Teórica e Trabalhos Relacionados

As Tecnologias Digitais de Informação e Comunicação (TDICs) ampliam as possibilidades de mediação pedagógica entre professores e estudantes, favorecendo a construção do conhecimento por meio da interação com ferramentas tecnológicas de apoio [Arôso Mendes Barbosa 2012].

A teoria sociointeracionista de *Vygotsky* sustenta que a aprendizagem ocorre por meio de interações sociais mediadas, nas quais a linguagem desempenha papel central na construção e internalização do conhecimento [Vygotsky 1978]. Nesse contexto, destaca-se a ZDP, definida como o intervalo entre o que o estudante consegue realizar de forma independente e aquilo que pode alcançar com o auxílio de um mediador mais experiente [Vygotsky 2008]. Ferramentas baseadas em LLMs, ao simularem diálogos educacionais e fornecerem *feedback* imediato, exemplos contextualizados e explicações em diferentes níveis de complexidade, podem atuar como mediadoras do processo de aprendizagem, apoiando estudantes em tarefas complexas, como a resolução de problemas de programação [Siqueira et al. 2025, Maia and Sarkis 2025].

Nesse cenário, interações sintéticas mediadas por LLMs ampliam a dimensão dialógica da aprendizagem no ambiente digital. Estudos recentes indicam que sistemas conversacionais baseados em IA podem favorecer ambientes de aprendizagem mais interativos, estimulando o engajamento do estudante e contribuindo para a construção do conhecimento por meio do diálogo com a máquina [Reis 2025]. Assim, ao simular práticas de interação educacional, essas tecnologias podem ampliar a autonomia discente mantendo coerência com os princípios da perspectiva sociointeracionista.

Diversos estudos recentes investigam o uso de IA no contexto educacional. Um trabalho brasileiro explorou o modelo Sabiá 2.0 em um assistente de ensino,

[5]https://www.flavioizo.com.br/pesquisas/syntheticConversation/2025a

demonstrando bom desempenho em tarefas de suporte educacional e maior aceitação entre estudantes devido à familiaridade linguística [Siqueira et al. 2025]. Outras pesquisas relatam a aplicação de LLMs no ensino de programação, evidenciando ganhos em engajamento, apoio pedagógico e geração de código [Maia and Sarkis 2025, Silva et al. 2024]. Estudos também destacam o potencial de *chatbots* educacionais como tutores virtuais disponíveis continuamente, capazes de reduzir a sobrecarga docente e oferecer *feedback* imediato aos estudantes [Alves et al. 2021, Cardoso et al. 2023]. Revisões da literatura reforçam benefícios como personalização da aprendizagem e automação de processos educacionais, mas apontam desafios relacionados à formação docente e à infraestrutura tecnológica [Barbosa 2023].

Apesar desses avanços, observa-se uma lacuna na literatura quanto à análise de conversas sintéticas entre professor e aluno, bem como à definição de métricas multidimensionais para avaliar a qualidade dessas interações. Trabalhos em outras áreas já exploram métricas para avaliação de diálogos sintéticos, como em interações paciente-médico [Haider et al. 2025], indicando a relevância desse tipo de abordagem.

Neste contexto, este trabalho investiga a utilização de interações sintéticas baseadas em LLMs para apoiar o ensino de programação, analisando a qualidade dessas conversas em cenários educacionais. Diferentemente de estudos que focam apenas no uso de chatbots como ferramentas de apoio, esta pesquisa busca compreender como essas interações podem ser estruturadas e avaliadas para favorecer o aprendizado dos estudantes. Como resultado, os achados podem subsidiar o desenvolvimento de materiais educacionais complementares e sistemas de Perguntas e Respostas (*Questions & Answers* – Q&A) mais eficazes para o ambiente educacional.

3. Metodologia

A pesquisa foi conduzida por um professor doutor que atua na área de programação há 21 anos. Os outros autores são alunos do curso de SI e estão regularmente matriculados nos 3º e 5º períodos do curso, ou seja, todos têm experiência de estudos com programação.

Para realizar os experimentos, foi necessário decidir sobre alguns pontos:

1. Quais LLMs utilizar nos experimentos?
2. Quais tópicos (conteúdos) dos roteiros de programação seriam interessante?
3. Como regidir um *prompt* padronizado e adequado para enviar à LLM?
4. Seleção dos critérios de avaliação dos roteiros
5. Procedimentos para avaliação dos roteiros, seguindo os critérios do passo 4
6. Escolha das métricas para avaliação do resultado dos experimentos

3.1. Seleção das LLMs

Após revisão exploratória da literatura e discussões entre os pesquisadores, foram listadas algumas LLMs que poderiam colaborar com este estudo. Após ter a lista elaborada, o próximo passo foi escolher quais utilizar, incluindo as devidas versões.

Assim, escolhemos quatro LLMs para efetuar os experimentos. A escolha de cada uma delas se deve por serem bem populares entre os usuários [Zhang et al. 2025]. As LLMs escolhidas foram: ChatGPT 5, Claude 4 Sonnet, Gemini 2.5 Pro e DeepSeek R1. A *DeepSeek* foi a representante *OpenSource*.

3.2. Seleção dos tópicos dos roteiros e execução dos *prompts*

Para a seleção dos tópicos dos roteiros, foi realizada uma reunião entre os pesquisadores na qual se decidiu solicitar às LLMs sugestões de roteiros que pudessem orientar esta pesquisa. As sugestões geradas foram submetidas a análise qualitativa pelos autores, considerando critérios como a frequência de dúvidas observadas entre estudantes de programação e a relevância dos temas para o processo de aprendizagem. Alguns tópicos foram unificados ou ajustados durante essa etapa de análise, resultando na definição dos 10 roteiros apresentados na Tabela 1, que foram utilizados neste experimento.

Tabela 1. Categorização dos Tópicos de Programação

#	Tópico
1	Sintaxe, Variáveis, Constantes e Tipos de Dados
2	Entrada e saída de dados
3	Estrutura Condicional e *Operadores Lógicos* (E, OU, NÃO)
4	Estrutura de Repetição
5	Operações matemáticas básicas
6	*Vetores/Arrays*
7	Modularização
8	Manipulação de *string* (concatenar e separar palavras)
9	Algoritmos de Ordenação
10	Tratamento de erros

Tendo os tópicos dos roteiros definidos, iniciamos, de maneira colaborativa, a criação de um *prompt* que fosse utilizado de maneira padronizada em todas as LLMs deste experimento. O texto padronizado pode ser visualizado a seguir.

Texto Padronizado Utilizado no *Prompt*

Olá _____ (nome da ferramenta). Eu gostaria que você criasse um diálogo entre um aluno e um professor de programação do curso de Sistemas de Informação. O aluno está no 1° período e tem dúvidas sobre um tópico específico relacionado à lógica de programação utilizando Portugol (um pseudocódigo utilizado para ensinar lógica de programação e algoritmos de maneira mais compreensível). O diálogo deve simular uma conversa real, permitindo que o aluno tire sua dúvida e aprenda o conteúdo, como acontece em sala de aula. Abaixo estão os detalhes da dúvida do aluno:
Dúvida: _____

Não informamos o contexto mais específico da dúvida para que a própria IA fosse criativa e tentasse alocar um exemplo pertinente (o que também faz parte da avaliação). Os exemplos foram executados nas LLMs em conversas separadas para não criar nenhum tipo de vício sequencial nas respostas.

3.3. Seleção dos critérios de avaliação dos roteiros

Inicialmente, fizemos uma reunião para pensar em quais critérios de avaliação poderiam ser utilizados. Nesse primeiro momento surgiram 5 critérios: *Ortografia*, *Qualidade Técnica*, *Relevância*, *Empatia* e *Conteúdo Realista*. Em seguida, conversamos com alguns profissionais da área pedagógica e estes solicitaram a inclusão de três critérios: *Adequação e Motodologia Pedagógica*, *Papel Hierárquico no Diálogo* e *Aplicabilidade*.

Posteriormente pesquisou-se em artigos científicos para corroborar a proposta e também encontrar um *baseline* próximo a proposta desta pesquisa. Após análise de um trabalho [Haider et al. 2025], adequamos a nomenclatura de alguns critérios. *Qualidade*

Técnica para *Precisão Técnica*, *Conteúdo Realista* para *Realismo* e *Papel Hierárquico no Diálogo* para *Consistência de Persona*. Os critérios adotados estão representadas na Tabela 2. Esses critérios geraram o índice *SCEPA*, que será melhor explicado na Seção 4.1.

Tabela 2. Critérios de Avaliação da Conversas Sintéticas

Critério	O que avaliar?
1. Precisão Técnica	Correção conceitual, Uso preciso de termos técnicos.
2. Adequação e Metodologia Pedagógica	Estratégias didáticas, Estímulo ao pensamento crítico, Clareza na explicação, Conexão teoria-prática.
3. Realismo	Naturalidade do diálogo, Linguagem apropriada, *Timing* das interações.
4. Empatia	Reconhecimento das dificuldades, Linguagem acolhedora, Paciência pedagógica, Motivação.
5. Consistência da Persona	Professor e aluno mantêm características, Coerência no conhecimento, Estilo comunicativo, Personalidade estável.
6. Relevância	Foco no problema, Informações úteis ao contexto do aluno.
7. Aplicabilidade	Aplicabilidade das respostas pelo aluno, Facilidade de implementação prática, Adequação ao nível do estudante.
8. Ortografia	Verificação de erros ortográficos na conversa.

Cada critério foi avaliado utilizando escala *Likert* de 5 pontos ordinais, com progressão qualitativa de "Inadequado" a "Excelente". A escala foi estruturada da seguinte forma: **5 - Excelente** (atende completamente ao critério com qualidade excepcional), **4 - Bom** (atende bem ao critério com pequenas limitações), **3 - Satisfatório** (atende adequadamente ao critério com qualidade média), **2 - Insuficiente** (atende parcialmente ao critério, necessitando melhorias), e **1 - Inadequado** (não atende ao critério, apresentando qualidade muito baixa). Para análise estatística, consideraram-se intervalos equivalentes entre os pontos da escala *Likert*, abordagem comum na avaliação de qualidade em contextos educacionais e tecnológicos [Sullivan and Artino Jr 2013].

3.4. Organização e aplicação dos questionários de avaliação

Seis estudantes do curso de SI foram selecionados como avaliadores, por estarem diretamente inseridos no contexto educacional analisado e familiarizados com os conteúdos de programação abordados nos roteiros. O total de avaliações foi calculado da seguinte forma: 10 roteiros × 4 LLMs × 6 avaliadores, totalizando 240 avaliações distintas.

Para viabilizar o processo avaliativo, foi desenvolvido um sistema *web* utilizando PHP e MySQL, que permitiu aos avaliadores realizarem suas análises de forma anônima e sem conhecimento prévio de qual LLM havia gerado cada conversa avaliada. Essa abordagem de avaliação cega (*blind evaluation*) assegura que os julgamentos sejam baseados exclusivamente na qualidade do conteúdo apresentado. Ao abrir o PDF, o avaliador tem acesso ao diálogo completo, podendo então avaliar segundo os critérios estabelecidos. Cada avaliador analisou individualmente as 40 conversas considerando 8 critérios de avaliação, gerando um total de 1920 registros de avaliações.

3.5. Métricas de Avaliação dos resultados

Para avaliar a confiabilidade e a qualidade dos dados coletados, bem como comparar o desempenho dos diferentes LLMs, foram empregadas algumas métricas estatísticas, cada uma com propósito específico na análise dos resultados.

3.5.1. Confiabilidade Inter-Avaliadores

Para mensurar a concordância entre os avaliadores, utilizou-se o coeficiente AC1 de Gwet [Gwet 2008]. Esta escolha metodológica justifica-se pela abordagem do AC1 em cenários onde possa ocorrer a prevalência extrema e distribuições marginais desbalanceadas [Gwet 2008, Wongpakaran et al. 2013], situação observada em alguns critérios desta pesquisa, especialmente nos critérios de *Relevância* e *Ortografia*, onde houve alta uniformidade de respostas.

O coeficiente AC1 calcula a probabilidade de concordância aleatória de forma menos sensível à distribuição das categorias de resposta. Valores próximos de 1 indicam maior concordância entre os avaliadores, enquanto valores próximos de 0 sugerem concordância equivalente ao acaso.

3.5.2. Estatísticas Descritivas

Para caracterizar o comportamento de cada critério do índice *SCEPA*, foram calculadas estatísticas descritivas incluindo média, desvio-padrão (DP), coeficiente de variação (CV), valores mínimo e máximo. O coeficiente de variação, expresso em percentual, permite avaliar a dispersão relativa dos dados, facilitando a comparação entre critérios com diferentes magnitudes, onde CVs elevados indicam maior heterogeneidade nas avaliações, sugerindo possível ambiguidade na interpretação do critério pelos avaliadores.

4. Resultados e Discussão

Esta Seção apresenta os resultados obtidos através de uma abordagem estruturada em três etapas principais. Inicialmente, realizou-se uma pesquisa exploratória com estudantes de Sistemas de Informação para validar a viabilidade e aceitação das conversas sintéticas como recurso educacional. Uma pesquisa preliminar foi realizada com 71 estudantes de Sistemas de Informação para avaliar a utilidade percebida de conversas sintéticas no ensino de programação. Os resultados indicam uma aceitação positiva, com 93% a concordar que tais diálogos poderiam apoiar a aprendizagem. Quanto ao engajamento, 53,5% dos estudantes demonstrou expectativa positiva e 81,7% reportaram que se sentiriam confortáveis em utilizá-los como recurso complementar de aprendizagem.

Em seguida, desenvolveu-se o *framework* aplicando o índice *SCEPA* para operacionalizar a avaliação das conversas, seguido da análise empírica, que envolveu 240 avaliações realizadas por 6 avaliadores da área de programação, abrangendo 40 conversas sintéticas distribuídas entre 10 cenários educacionais de programação, geradas por quatro LLMs distintos.

4.1. Framework *SCEPA*: Implementação e Validação

Como resultado desta pesquisa, foi desenvolvido um *framework* configurável que implementa o índice *SCEPA* para avaliação de conversas sintéticas em contextos educacionais. O sistema integra os 8 critérios definidos pelo índice, permitindo ajustar os pesos de cada critério conforme o contexto pedagógico analisado.

O protótipo *web* permite o cadastro de conversas sintéticas geradas por diferentes LLMs e a realização de avaliações por especialistas. Os avaliadores analisam cada conversa utilizando os 8 critérios definidos pelo índice, por meio de escalas de 5 pontos de *Likert*, garantindo consistência no processo de avaliação.

O protótipo implementa a fórmula (1) do índice *SCEPA* como média ponderada configurável:

$$SCEPA_{score} = \sum_{i=1}^{n} w_i \cdot C_i \quad (1)$$

onde n representa a quantidade de critérios avaliados (8), w_i representa o peso configurável do critério i e C_i o *score* normalizado do critério i. É importante ressaltar que para este estudo, utilizamos pesos equitativos ($w_i = 0.125$).

O sistema também disponibiliza uma interface para visualização e análise dos resultados, permitindo a configuração de pesos para cada critério e o recálculo automático do índice *SCEPA*. Essa funcionalidade possibilita adaptar a análise a diferentes objetivos pedagógicos e apoiar a seleção de conversas adequadas para atividades educacionais.

4.2. Confiabilidade Inter-Avaliadores

Na literatura ainda não há consenso sobre a utilização de escalas de valores para os índices de *AC1 de Gwet*. Alguns autores utilizam as escalas de *Landis and Roch* [Landis and Koch 1977] para categorizar o resultado do índice [Wongpakaran et al. 2013] e outros não indicam o seu uso, alegando que só podem ser utilizados para o índice de Kappa [Vach and Gerke 2023]. Diante da ausência de consenso na literatura quanto à categorização dos valores do AC1, optou-se por apresentar os índices obtidos de forma descritiva, permitindo a interpretação direta dos níveis de concordância observados. O resultado referente a concordância entre os avaliadores são apresentados na Tabela 3.

Tabela 3. Índices de Confiabilidade Inter-Avaliadores por Critério

Critério	Precisão Técnica	Adequação Pedagógica	Realismo	Empatia	Consistência de Persona	Relevância	Aplicabilidade	Ortografia
AC1 de Gwet	0,5933	0,2062	0,2625	0,3125	0,3854	0,6225	0,2604	0,7417
Média Geral	**0,4231**							

Os resultados evidenciam elevada concordância nos critérios *Ortografia* (0,7417), *Relevância* (0,6225) e *Precisão Técnica* (0,5933), sugerindo consistência na avaliação desses aspectos. No entanto, critérios como *Adequação Pedagógica* (0,2062), *Aplicabilidade* (0,2604) e *Realismo* (0,2625) apresentaram valores inferiores, indicando maior variabilidade entre os avaliadores. O índice médio geral de 0,4231 situa-se na faixa de confiabilidade moderada, demonstrando que, de forma global, houve consistência aceitável, embora alguns critérios demandem maior refinamento conceitual ou instruções mais precisas para reduzir discrepâncias e aprimorar a reprodutibilidade das avaliações.

Essa discrepância entre os critérios de maior e menor confiabilidade pode estar relacionada ao grau de objetividade inerente a cada critério avaliado. Aspectos mais concretos, como ortografia, relevância e precisão técnica, favorecem maior consistência entre avaliadores. Por outro lado, critérios de natureza mais interpretativa, como adequação pedagógica, realismo e aplicabilidade, tendem a gerar diferentes entendimentos, refletindo a subjetividade envolvida nesses julgamentos. Esse padrão sugere a necessidade de aprimorar as definições operacionais e fornecer exemplos mais claros para reduzir divergências nos critérios menos objetivos.

4.3. Estatísticas Descritivas dos Critérios SCEPA

A análise descritiva dos 8 critérios do índice *SCEPA* revelou variações tanto em termos de médias quanto de dispersão das avaliações, conforme podemos analisar na Tabela 4.

Tabela 4. Estatísticas Descritivas dos Critérios do Índice *SCEPA* (n=240)

Critério	Média	DP	CV(%)	Min	Max
Precisão Técnica	4,770	0,542	11,375	2,0	5,0
Adequação Pedagógica	4,108	1,041	25,347	1,0	5,0
Realismo	3,895	1,310	33,635	1,0	5,0
Empatia	4,195	1,112	26,514	1,0	5,0
Consistência de Persona	4,458	0,918	20,599	1,0	5,0
Relevância	4,850	0,401	8,286	3,0	5,0
Aplicabilidade	4,270	0,953	22,331	1,0	5,0
Ortografia	4,829	0,563	11,676	1,0	5,0

A Tabela 4 sintetiza as estatísticas descritivas (n = 240) para os 8 critérios avaliativos. Em termos gerais, as médias situam-se em faixa elevada da escala (variam de 3,895 a 4,850 numa escala de *Likert* de 5 pontos), o que indica uma avaliação globalmente positiva dos itens mensurados. Entretanto, a análise conjunta da média com as medidas de dispersão (desvio-padrão e coeficiente de variação) revela diferenças quanto consenso entre os respondentes, informação crucial para decisões de revisão do instrumento.

Os dados evidenciam heterogeneidade entre os critérios. *Relevância* apresentou a média mais elevada (4,850) e o menor coeficiente de variação (CV = 8,28%), sugerindo elevada concordância e possível efeito teto (valores próximos do máximo). Em contraste, critérios como *Realismo* (Média = 3,895; CV = 33,63%), *Empatia* (Média = 4,195; CV = 26,54%), *Consistência de Persona* (Média = 4,458; CV = 20,59%) e *aplicabilidade* (Média = 4,270; CV = 22,31%) exibiram maior valor de dispersão, refletindo interpretações mais divergentes entre os avaliadores. *Adequação Pedagógica* (Média = 4,108; CV = 15,27%) ocupou uma posição intermediária, com variabilidade moderada. Esses achados sugerem boa aceitação geral das conversas sintéticas, porém apontam a necessidade de aprimorar as definições operacionais e os procedimentos de calibração avaliativa (dos avaliadores), a fim de reduzir ambiguidades em critérios mais subjetivos. Essa análise confirma o que já havia sido identificado na Seção 4.2.

4.4. Performance dos *Large Language Models*

A Tabela 5 apresenta o *ranking* das quatro LLMs avaliadas pelo índice *SCEPA*. Observa-se liderança do *Gemini*, seguido por *DeepSeek* e *Claude*, enquanto o *ChatGPT* apresentou desempenho inferior. A análise indica que *Gemini* e *DeepSeek* obtiveram desempenhos quase equivalentes, próximos ao limite superior do índice *SCEPA*, com o *Claude* também posicionado nessa faixa, ainda que ligeiramente abaixo, sugerindo que esses três modelos apresentam desempenho competitivo no contexto avaliado.

Tabela 5. Ranking dos LLMs segundo o índice *SCEPA*

Posição	LLM	Índice *SCEPA*	IC 95%	N Avaliações
1º	Gemini 2.5 Pro	4,714	4,630 - 4,798	60
2º	DeepSeek R1	4,631	4,534 - 4,728	60
3º	Claude Sonnet 4	4,552	4,424 - 4,679	60
4º	ChatGPT-5	3,791	3,624 - 3,959	60

Em contraste, o *ChatGPT-5* obteve pontuação menor, com um *gap* de 0,961 pontos (19,22% da escala) em relação ao melhor modelo. Esses resultados indicam que, enquanto *Gemini* e *DeepSeek* despontam como escolhas preferenciais e *Claude* pode atuar como alternativa intermediária, o *ChatGPT-5* apresenta limitações na geração de conversas

sintéticas, demandando aprimoramentos para alcançar desempenho competitivo.

Chegamos a essa conclusão em relação ao *ChatGPT-5* após análise conjunta dos roteiros com os avaliadores. Apesar da boa precisão teórica, as respostas mostraram-se excessivamente mecânicas e pouco didáticas, assemelhando-se a uma *receita de bolo*: um passo a passo objetivo para resolver a dúvida do aluno, sem priorizar o aspecto central da interação pedagógica, a relação aluno–professor.

Além disso, um avaliador destacou é possível identificar os roteiros gerados pelo *ChatGPT-5*, devido ao padrão rígido em etapas e pela objetividade das explicações, características marcantes desse modelo. Em contraste, modelos como o *Gemini Pro* e o *DeepSeek R1* se sobressaíram, pois conseguiram simular de forma mais natural a interação aluno-professor, demonstrando maior domínio de estratégias educacionais e transmitindo maior humanidade nos roteiros, sem comprometer a clareza e a precisão do conteúdo.

4.5. Limitações Metodológicas

Esta pesquisa apresenta algumas limitações. Primeiramente, a confiabilidade inter-avaliadores foi moderada (AC1 = 0,4231), refletindo a natureza subjetiva de parte dos critérios avaliados. Além disso, o estudo foi conduzido no contexto de dúvidas de programação em cursos de SI, o que pode limitar a generalização dos resultados para outros domínios educacionais. Embora a amostra de 240 avaliações seja adequada para a análise realizada, estudos futuros podem ampliar o número de avaliadores e cenários analisados. Por fim, os resultados refletem o desempenho das LLMs no período em que o estudo foi conduzido, sendo necessárias reavaliações diante evolução desses modelos.

5. Considerações Finais

Este trabalho apresentou o índice *SCEPA* e um *framework* para avaliação de conversas sintéticas geradas por LLMs no ensino de programação. Os resultados indicam que esse tipo de material possui potencial como recurso educacional complementar, com alta aceitação pelos estudantes seja pela utilidade pedagógica, seja pelo conforto de uso.

A análise comparativa entre quatro LLMs mostrou diferenças relevantes na qualidade pedagógica das conversas geradas. O *Gemini 2.5 Pro* apresentou o melhor desempenho geral, seguido por *DeepSeek R1* e *Claude Sonnet 4*, enquanto o *ChatGPT-5* obteve pontuação inferior. Os resultados também indicam maior concordância entre avaliadores em critérios objetivos (como ortografia e relevância) e maior variabilidade em critérios interpretativos, como adequação pedagógica e aplicabilidade.

Como principais contribuições, destacam-se: (i) a proposição e avaliação empírica do índice *SCEPA* para análise de conversas educacionais geradas por IA; (ii) um *framework* configurável que permite adaptar pesos dos critérios conforme o contexto educacional; e (iii) um conjunto de 40 conversas avaliadas por múltiplos avaliadores, disponibilizado para apoiar futuras pesquisas em IA aplicada à educação.

Como trabalhos futuros, pretende-se validar a abordagem em contextos reais de sala de aula, investigar o impacto das conversas sintéticas no desempenho dos estudantes e expandir a aplicação do índice *SCEPA* para outros domínios educacionais.

6. Declaração sobre uso de Inteligência Artificial

Em conformidade com as diretrizes da SBC, declaramos o uso de ferramentas de Inteligência Artificial generativa nesta pesquisa. A ferramenta Claude 4.5 Sonnet foi utilizada como apoio na revisão linguística do texto e em ajustes pontuais de estilo, sem alteração do conteúdo conceitual elaborado pelos autores. Também foi utilizada como auxílio na elaboração do *layout* CSS da interface do *framework*. As conversas sintéticas analisadas neste estudo foram geradas pelas LLMs ChatGPT 5, Claude 4 Sonnet, Gemini 2.5 Pro e DeepSeek R1, que constituem o objeto da investigação, sendo posteriormente avaliadas por avaliadores humanos.

Referências

Alves, R. d. S., Nascimento, G. M. d., and Sousa, R. R. d. (2021). Elementos do Emprego de Chatbots para Auxílio no Ensino de Programação: Uma Revisão Sistemática da Literatura. *Brazilian Journal of Development*, 7(5):43908–43927.

Arôso Mendes Barbosa, C. M. (2012). A Aprendizagem Mediada por TIC: Interação e Cognição em Perspectiva. *Revista Brasileira De Aprendizagem Aberta e a Distância*, 11.

Barbosa, C. R. d. A. C. (2023). Transformações no Ensino-Aprendizagem com o Uso da Inteligência Artificial: Revisão Sistemática da Literatura. *RECIMA21 - Revista Científica Multidisciplinar*, 4(3).

Bortolazzo, S. F. (2024). Storytelling: Entre Usos, Benefícios e Aprendizagens. *Ensino em Re-Vista*, 31(Contínua):1–24.

Cardoso, F. S., Pereira, N. d. S., Braggion, R. C., Chaves, P., and Andrioli, M. (2023). O Uso da Inteligência Artificial na Educação e Seus Benefícios: Uma Revisão Exploratória e Bibliográfica. *Revista Ciência em Evidência*, 4.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Haider, S. A., Prabha, S., Gomez-Cabello, C. A., Borna, S., Genovese, A., Trabilsy, M., Collaco, B. G., Wood, N. G., Bagaria, S., Tao, C., and Forte, A. J. (2025). Synthetic Patient-Physician Conversations Simulated by Large Language Models: A Multi-Dimensional Evaluation. *Sensors*, 25(14):4305.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

M. Valença, M. and Balthazar Tostes, A. P. (2019). O Storytelling como Ferramenta de Aprendizado Ativo. *Carta Internacional*, 14(2).

Maia, S. M. and Sarkis, L. C. (2025). Utilização de LLM como Ferramenta de Apoio no Ensino-Aprendizagem de Programação Python para Iniciantes: Um Relato de Experiência. In *Workshop sobre Educação em Computação (WEI)*, pages 385–396. SBC.

Reis, C. d. S. (2025). Inteligência Artificial Generativa, Metodologias Ativas e Escolarização Aberta: Desafios e Potencialidades no Ambiente Educacional no Ensino Superior. Dissertação de mestrado, Universidade Federal de Santa Catarina, Florianópolis, SC.

Silva, T. L. d., Vidotto, K. N. S., Tarouco, L. M. R., and Silva, P. F. d. (2024). Inteligência Artificial Generativa no Ensino de Programação: Um Mapeamento Sistemático da Literatura. *Revista Novas Tecnologias na Educação*, 22(1):262–272.

Siqueira, E., Portela, C., and Moraes, A. (2025). Teaching Assistant Based on a Brazilian Large Language Model. In *Anais do XXI Simpósio Brasileiro de Sistemas de Informação*, pages 300–308, Porto Alegre, RS, Brasil. SBC.

Sociedade Brasileira de Computação (2025). *Grandes Desafios da Computação no Brasil 2025–2035*. Sociedade Brasileira de Computação, Porto Alegre, Brasil.

Sullivan, G. M. and Artino Jr, A. R. (2013). Analyzing and interpreting data from likert-type scales. *Journal of fvachgraduate medical education*, 5(4):541.

Vach, W. and Gerke, O. (2023). Gwet's AC1 is not a Substitute for Cohen's Kappa - A Comparison of Basic Properties. *MethodsX*, 10:102212.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Vygotsky, L. S. (2008). *Pensamento e Linguagem*. Martins Fontes, São Paulo, 4 edition.

Wongpakaran, N., Wongpakaran, T., Wedding, D., and Gwet, K. L. (2013). A Comparison of Cohen's Kappa and Gwet's AC1 When Calculating Inter-Rater Reliability Coefficients: a Study Conducted with Personality Disorder Samples. *BMC Medical Research Methodology*, 13(1):61.

Zhang, Y., Chen, T. Y., Huang, R., Pike, M., Towey, D., Ying, Z., and Zhou, Z. Q. (2025). Comparative Analysis of Styles in LLM-Generated Code for LeetCode Problems: A Preliminary Study. In *2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1625–1630.