Revista de Estudos Interdisciplinares

ENGENHARIA DE *SOFTWARE* GENERATIVA/DETERMINÍSTICA: PEDAGOGIA COM SISTEMAS MULTIAGENTES LOCAIS GERENCIANDO CARGAS COGNITIVAS/COMPUTACIONAIS

GENERATIVE/DETERMINISTIC SOFTWARE ENGINEERING: PEDAGOGY WITH LOCAL MULTI-AGENT SYSTEMS MANAGING COGNITIVE/COMPUTATIONAL LOADS

Recebido em: 05/05/25
Aceito em: 30/04/2026
Publicado em: 31/05/2026

Hélio Craveiro Pessoa Júnior[1]
Universidade de Brasília

Resumo: A integração de sistemas inteligentes desafia a educação em desenvolvimento de *software*. Políticas atuais negligenciam a complexidade de tecnologias emergentes, como o ajuste fino de modelos com baixo uso de memória e a orquestração de múltiplos modelos inteligentes locais, além das limitações significativas de recursos computacionais. Este estudo propõe um modelo pedagógico centrado na interdependência crítica entre o esforço mental do estudante e a demanda de processamento da máquina. Achados empíricos indicam que otimizar a demanda de processamento (ex: com quantização agressiva) pode elevar o esforço mental, e a complexidade de sistemas com múltiplos modelos inteligentes amplifica ambas as cargas. O modelo busca um "ponto ideal" pedagógico, onde uma demanda de processamento moderada (ex: quantização mais estável) minimiza o esforço mental, com adaptação à experiência do aluno. Visa desenvolver metacognição para gerenciar esses compromissos em tarefas complexas, especialmente com *hardware* restrito. Estratégias didáticas incluem sequenciamento progressivo, atividades comparativas e execução assíncrona para alta demanda de processamento, com resultados gerenciados reflexivamente. Ferramentas de apoio, como a plataforma *CCLOLMAS* (disponível em https://github.com/cclolmas/app), tornam essa dinâmica visível. O objetivo é reformular currículos, alinhando a formação às demandas da indústria e capacitando estudantes a gerenciar efetivamente esses compromissos essenciais na prática profissional.

Palavras-chave: Educação em Engenharia de *software*; Sistemas Multiagente Locais; Carga Cognitiva; Carga Computacional; Estratégias Pedagógicas.

Abstract: The integration of intelligent systems challenges software development education. Current policies overlook the complexity of emerging technologies, such as low-memory fine-tuning of models and the orchestration of multiple local intelligent models, alongside significant computational resource limitations. This study proposes a pedagogical framework centered on the critical interdependence between student mental effort and machine processing demand. Empirical findings indicate that optimizing processing demand (e.g., with aggressive quantization) can elevate mental effort, and the complexity of systems with multiple intelligent models amplifies both burdens. The framework seeks a pedagogical "sweet spot," where moderate processing demand (e.g., more stable quantization) minimizes mental effort, with adaptation to the student's experience. It aims to develop metacognition for managing these trade-offs in complex tasks, especially with restricted hardware. Didactic strategies include progressive sequencing, comparative activities, and asynchronous execution for high processing demand, with results managed reflectively. Support tools, such as the CCLOLMAS platform (available at https://github.com/cclolmas/app), make this dynamic visible. The objective is to reformulate curricula, aligning training with industry demands and empowering students to effectively manage these essential trade-offs in professional practice.

Keywords: Software Engineering Education; Local Multi-Agent Systems; Cognitive Load; Computational Load; Pedagogical Strategies.

[1] Aluno do Programa de Pós-Graduação em Educação – Modalidade Profissional da Universidade de Brasília (UnB). Brasil, Distrito Federal, Brasília. E-mail: pessoajr@gmail.com

INTRODUÇÃO

A oferta de vagas para Engenharia de *software* em instituições federais brasileiras varia de 30 a 176. UTFPR lidera (176 vagas), seguida por UFC (150) e UnB (112). UFMS (70), UFG, UFERSA e UFRN (60 cada uma), UNIPAMPA e UFAM (50 cada uma), IFAM, IFSP e IFPR (40 cada uma), IFG (35) e IFPE (30) completam o quadro, refletindo demanda.

A consolidação do BRICS-Educação (Cúpula de Kazan, 2024) adquire relevância estratégica para a formação em *SE* no Brasil. Este fórum, que culminará na Reunião Ministerial de 5 de junho, preconiza: (1) educação digital com arquiteturas de *AI* local; (2) políticas para integração ética de sistemas multiagentes; (3) harmonização de diretrizes para ensino superior em tecnologias emergentes. A convergência BRICS e necessidades nacionais (Brasil, 2019, 2025) materializa-se no programa "Escolas Conectadas" do MEC, alinhado às recomendações do bloco. Essa sinergia estrutura propostas pedagógicas, especialmente na gestão das cargas cognitiva e computacional em ambientes de recursos limitados.

A Engenharia de *software* (*SE*) é transformada pela *AI* (Hassan *et al.*, 2024), impactando todo o ciclo de desenvolvimento. Rodar modelos como *Qwen* localmente é viável com técnicas como quantização e formatos como GGUF, permitindo *LMASs* (Wu *et al.*, 2023) em *hardware* comum, com vantagens de custo, latência e privacidade (Husom *et al.*, 2025; Jia *et al.*, 2025; Birkmose *et al.*, 2025).

Essa mudança coincide com discussões no Brasil sobre *AI* na educação (defesa da Associação Brasileira das Instituições Comunitárias de Educação Superior por uma maior integração com o setor produtivo, por ex.) e sociedade (interesse em *AI*/Ciência de Dados). Contudo, a educação em *SE* no Brasil (Brasil, 2014, 2016, 2019) dificilmente incorpora as competências em *AI*.

Este artigo foca na interdependência entre Carga Cognitiva (*CL*) humana (Sweller *et al.*, 1998) e Carga Computacional (*CompL*) da máquina em tarefas de *AI* complexas (ex.: ajuste fino via *QLoRA*, Dettmers *et al.*, 2023; orquestração de *LMASs* - Luo *et al.*, 2025; Zhang *et al.*, 2025). Equilibrar cargas é vital ao aprendizado (Retzlaff *et al.*, 2024), com um "ponto ideal" (H3) onde *CompL* moderada minimiza *CL*. Tarefas como *QLoRA* ou *LMASs* geram alta *CL* e *CompL* em ambientes acadêmicos com recursos limitados (Bai *et al.*, 2024; Ni *et al.*, 2025), dificultando o aprendizado.

Propõe-se uma abordagem pedagógica para gerenciar a interação entre Cognição e Computação (*CL-CompL*): (1) Reconhecer o balanço *CL-CompL* (ex.: quantização); (2) Ferramentas para monitorar *CL* subjetiva (Paas *et al.*, 2003) e *CompL* objetiva, promovendo

metacognição (Zimmerman, 2002); (3) Adaptar estratégias à expertise, gerenciando complexidade para evitar sobrecarga e efeito de reversão da expertise (Kalyuga *et al.*, 2003).

O objetivo é guiar alunos ao "ponto ideal" (H3) de eficiência cognitiva e computacional. A estrutura visa preencher lacunas na educação em SE, integrar técnicas como *QLoRA* e *LMASs*, e alinhar o ensino superior brasileiro às demandas da *AI* e demandas nacionais.

REVISÃO DA LITERATURA

A Teoria da Carga Cognitiva (*CLT*) (Sweller *et al.*, 1998) elucida como a elevada *ICL* de domínios complexos como *LMASs* e *QLoRA* interage com o *design* instrucional (determinando *ECL*) para modular a carga *CL-CompL* na aprendizagem. Os Saberes Digitais Docentes (SDD) recontextualizam a *CLT*: *CompL* local efetivamente modula *CL* total. O educador deve então discernir que: (1) Intervenções para mitigar *CompL* (ex.: Q4) podem exacerbar *CL* (Hipótese 1) devido a instabilidades e à complexidade na depuração (Pu *et al.*, 2025); (2) Gestão criteriosa da *CompL* (ex.: Q8, execução assíncrona) é uma estratégia pedagógica fundamental para regular *CL*, buscando o "ponto ótimo" cognitivo-computacional (Hipótese 3); (3) Assincronia desafia *CL*: latência pode induzir ansiedade (*ECL*), e a análise posterior de outputs complexos exige reativação de esquemas mentais (*GCL*); (4) SDD implicam projetar estrutura de apoio (Wood *et al.*, 1976) para gerenciar ambas as cargas, decompor tarefas (mitigando *ICL*), prover interfaces claras (minimizando *ECL*), e fomentar metacognição sobre o balanço *CL-CompL* (otimizando *GCL* - Zimmerman, 2002). O planejamento cuidadoso deve incluir atividades durante intervalos de processamento.

O Conhecimento Tecnológico, Pedagógico e de Conteúdo (*TPACK*) (Mishra e Koehler, 2006) apoia docente de *SE* ensinando *LMASs* localmente com recursos limitados. Aqui, SDD manifestam-se como *TPACK* específico, focado na gestão ativa da interação *CL-CompL*. Este *TPACK* especializado exige: (1) Conhecimento do Conteúdo (*C*) focado em aspectos de *AI* relevantes à prática local (nuances de *LMASs*, *fine tuning*, implicações Q4 vs. Q8, *LLMOps* em cenários restritos); (2) Conhecimento Pedagógico (*P*) aplicando *CLT* recontextualizada, metodologias ativas e design instrucional para assincronia e gestão de expectativas e latências; (3) Conhecimento Tecnológico (*T*) além da operação de ferramentas (*LangChain* - Luo *et al.*, 2025), incluindo compreensão do impacto da *CompL*, limitações (instabilidade Q4, depuração *LMASs*), e capacidade de instrumentar/visualizar dinâmica *CL-CompL*. Inclui configurar ambientes, selecionar modelos (ex.: Phi-4-mini, Mistral - Jiang *et al.*, 2023) compatíveis com *hardware* (*VRAM*) e gerenciar recursos. Competência fundamental (*TPACK*) está nas

interseções: conceber tarefas autênticas (*CK+PK+TK*) usando *LMASs* (*TK*) para problemas reais (*CK*), com uma pedagogia (*PK*) mediando balanço *CL-CompL* e assincronia, dentro das limitações de *hardware* (*TK*). Envolve selecionar/adaptar ferramentas (*TK*) que forneçam *feedback* sobre *CL*/*CompL* (*PK*), facilitando a autorregulação metacognitiva (*PK+CK*).

A Teoria da Atividade (*AT*) (Engestrom, 1987) e Cognição Distribuída (*dCog*) (Hutchins, 1995) modelam aprendizagem como um sistema sócio-técnico, capturando interação, especialmente com processos assíncronos. (1) Sistema de Atividade - Aprendizagem de *LMASs* envolve Sujeito (estudante), Ferramentas (*LLMs* locais, *frameworks*, *IDEs*, *hardware* com *CompL*), Objeto (aplicação de *LMASs*, gestão *CL-CompL*), Regras (pedagogia, avaliação, protocolos para assincronia), Comunidade (pares, docente) e Divisão do Trabalho (distribuição cognitiva aluno-sistema, esp. inferências longas); (2) Mediação e Cognição Distribuída - Artefatos tecnológicos (incl. *hardware* limitado) medeiam aprendizagem, cognição é distribuída entre estudante e ambiente. SDD envolvem *design* dessa distribuição (ex.: ferramentas externalizando métricas de *CompL* para auxiliar monitoramento); (3) Contradições como Força Motriz - Tensão entre aprendizagem profunda (alta *GCL* desejável) e limitações de recursos (alta *CompL*/*ECL* indesejada) é uma contradição central (*AT*), SDD eficazes tornam contradição explícita, usando-a como catalisador para competências metacognitivas (gerenciar o balanço *CL-CompL*); (4) Ciclos Assíncronos - *AT*/*dCog* modela ciclo completo, docente projeta sistema para manter engajamento, tarefas preparatórias antes da execução, análise/reflexão após, e *feedback* conectando fases, assegurando continuidade cognitiva.

SDD para professores de *SE* na era da *AI* local sob restrições e inferências assíncronas demandam uma pedagogia integradora, além da proficiência isolada. Requer uma competência ancorada na *CLT* recontextualizada (*CompL*, assincronia), informada por *TPACK* focado na interação humano-IA restrita, e por *AT*/*dCog* para *design* de sistemas sociotécnicos resilientes. Isso habilita educador a preparar engenheiros para orquestrar a interação cognição-computação em cenários com recursos limitados, como no contexto brasileiro. Esta orientação transmuta limitação em oportunidade, alinhando a práxis educacional às exigências da profissão.

A computação em tempo de espera, "*Sleep-time compute*" (Lin *et al.*, 2025) oferece um paralelo para dinâmica *CL-CompL*. Computações sobre contexto (c) antes da consulta (q), gerando uma representação enriquecida (c'), deslocam parte do esforço computacional do "*test-time*" para "*sleep-time*". Essa estratégia visa reduzir *CompL* na consulta, podendo aliviar *CL* do usuário final. Em aplicações "*stateful*", como agentes de codificação ou assistentes

conversacionais, essa abordagem ajuda a gerenciar *CompL* e equilibrar custos cognitivos e computacionais (H1, H2).

O *"Sleep-time compute*" introduz uma amortização do custo computacional. Com múltiplas consultas relacionadas (q1...qn) sobre mesmo contexto (c), custo da pré-computação para gerar c' pode ser diluído, reduzindo custo médio por consulta (Lin *et al.*, 2025). Essa otimização temporal é uma habilidade essencial para engenheiros de *SE* que estejam gerenciando sistemas de *AI* complexos. Contudo, a eficácia depende da "previsibilidade da consulta" – capacidade de antecipar interações futuras. Estudantes precisam desenvolver capacidade técnica e metacognição para avaliar balanços e alocar recursos computacionais (*sleep-time vs. test-time*), buscando o "ponto ideal" (H3) de eficiência e desempenho, indispensável para inovação responsável em *AI*.

Gestão de recursos em agentes de *LLM* estende-se à otimização de ferramentas externas na inferência (*framework OTC*, Wang *et al.*, 2025b), usando Aprendizado por Reforço *(RL*) para minimizar chamadas sem sacrificar precisão, abordando balanço *CompL*-eficácia. Quando os *LLMs* maiores são proibitivos (*CompL*), colaboração entre os *LLMs* menores emerge como uma alternativa. *Frameworks* como *GRA* (GAO *et al.*, 2025), com agentes especializados (Gerador, Revisor, Adjudicador) para síntese de dados, e *CORY* (Ma *et al.*, 2025), com coevolução pioneer-observer para *fine tuning* robusto, demonstram que sistemas multiagente podem atingir ou superar a qualidade de modelos monolíticos maiores.

Essa abordagem ressoa com princípios da cognição distribuída (*dCog*), onde inteligência é distribuída entre múltiplos agentes (humanos ou artificiais) e artefatos. Construção desses agentes colaborativos exige arquiteturas definidas, como *RAISE* (Liu *et al.*, 2025a), com uma memória dual para gerenciar contexto, e protocolos padronizados, como *MCP* (Ray, 2025), para interoperabilidade segura. Contudo, complexidade de sistemas multiagentes e inferências longas introduz desafios de robustez e falhas (Ferrag *et al.*, 2025), exigindo otimização e garantias formais de consistência, como via Teoria das Categorias (Ghosh *et al.*, 2025), para mitigar "reasoning buckling" em tarefas científicas ou de SE.

PROCEDIMENTOS METODOLÓGICOS

Metodologia visou criar um modelo pedagógico para integrar tópicos de *AI* emergentes (*LMASs*; *QLoRA*, Dettmers *et al.*, 2023) em currículos de SE. Uma abordagem sintetizou teoria e empiria, articulando dinâmica *CL-CompL* com teorias educacionais (*CLT* - Sweller *et al.*, 1998; Van Merrienboer; Sweller, 2005), *HCI* (Nielsen, 1993; Norman, 1988; Amershi *et al.*,

2019; Xu, 2019) e contexto brasileiro, como *hardware* modestos e as Diretrizes Curriculares Nacionais (DCNs). Processo integrou: (1) achados empíricos (H1-H4), com análise *CL-CompL* via modelagem estatística (GLMM) para complexidade e moderação por expertise; (2) *CLT*; (3) princípios de *HCI* (usabilidade, *design*, *feedback*); e (4) análise contextual. Buscou-se equilibrar teoria, evidência e viabilidade.

Desenvolvimento do modelo seguiu um processo iterativo: tradução de achados *CL-CompL* ao contexto educacional; mapeamento desafios a objetivos; aplicação de princípios da *CLT* (gerenciar *ICL*, minimizar *ECL*, otimizar *GCL* via metacognição - Zimmerman, 2002; *PBL*/*PjBL* - Hmelo-Silver, 2004; Thomas, 2000); incorporação restrições contextuais (modelos moderados, gestão de *VRAM*, execução assíncrona); e síntese em modelo coerente. Incluiu desenvolvimento e validação inicial (usabilidade, pilotos) da plataforma *CCLOLMAS* para operacionalizar estratégias.

A operacionalização pedagógica definiu *CL* como esforço mental (medido via *SEQ*/Paas *et al.*, 2003, observação, desempenho) e *CompL* como a respectiva demanda de recursos (medida objetivamente - *RAM*, *VRAM*). A gestão pedagógica da *CompL* incluiu um cuidadoso planejamento e sua integração curricular. O foco principal foi equilibrar *CL* e *CompL* para experiências desafiadoras, mas viáveis, promovendo aprendizado profundo e habilidades transferíveis em *SE* na complexa era da *AI*.

Esta análise aborda: (1) oportunidades/desafios da *AI*, especialmente de *LMAS*, na educação em *SE* (Siddiq *et al.*, 2024; Yetistiren *et al.*, 2024); (2) panorama de políticas educacionais brasileiras frente à formação tecnológica e em *AI*; (3) aplicação da *CLT* (Sweller *et al.*, 1998; Van Merrienboer; Sweller, 2005) para entender dificuldades em domínios complexos como *SE* e *AI* (Duran *et al.*, 2022; Xinogalos, 2021); e (4) desafios pedagógicos do ensino de ajuste fino local (*QLoRA*, Dettmers *et al.*, 2023) e orquestração de *LMASs*. Objetivo é identificar lacunas justificando um modelo pedagógico baseado na gestão *CL-CompL*, considerando restrições (Bai *et al.*, 2024).

A rápida evolução da SE, impulsionada pela GenAI, com ferramentas como *Code Llama* (Roziere *et al.*, 2023), *StarCoder* (Li *et al.*, 2023) e assistentes de codificação (Bird *et al.*, 2022; Kazemitaar *et al.*, 2024), promete ganhos de produtividade (Leinonen *et al.*, 2023) e oportunidades de focar em habilidades de alto nível, personalizar aprendizado (Kotalwar *et al.*, 2025) e desenvolver engenharia de *prompt* (Spiess *et al.*, 2025). Contudo, surgem desafios (Hassan *et al.*, 2024), tais como aprendizado superficial (Piccolo *et al.*, 2024), comprometimento de fundamentos (Luan *et al.*, 2023), e avaliação crítica das saídas instáveis

ou inseguras de modelos (Perry *et al.*, 2022). Interfaces e processos inadequados aumentam *ECL* (Sweller, 2010; Mayer; Moreno, 2003), e falta de atualização curricular/docente é uma lacuna (Siddiq *et al.*, 2024), especialmente para tópicos sofisticados como ajuste fino local (Ayupov; Chirkova, 2022; Liu *et al.*, 2023) e orquestração de *LMASs* (Wu *et al.*, 2023; Zhang *et al.*, 2025; OpenAI, 2025; Ghosh, 2025; Ahmadi *et al.*, 2024), áreas negligenciadas pela pesquisa focada em ferramentas básicas (Yetistiren *et al.*, 2024).

O sistema educacional brasileiro, via Plano Nacional de Educação (PNE) (Brasil, 2014; INEP, 2023) e DCNs, responde de forma bastante limitada à atual revolução da *AI*. O PNE atual não aborda especificamente *AI*, e as revisões (MEC, 2024) propõem integrar pensamento computacional (Denning; Tedre, 2019) apenas genericamente. As DCNs enfatizam metodologias ativas, mas carecem de diretrizes claras para conteúdos de ponta em *AI* ou para a gestão pedagógica de *CL* e *CompL*. Essa falta de especificidade reforça a necessidade de políticas mais direcionadas, considerando os contextos de desenvolvimento (Mhlanga, 2023).

CLT (Sweller *et al.*, 1998; Van Merrienboer; Sweller, 2005) distingue Carga Intrínseca (*ICL* - complexidade inerente), Carga Extrínseca (*ECL* - *design* instrucional) e Carga Germânica (*GCL* - construção de esquemas) (Sweller, 2010). Na educação em *SE* (Duran *et al.*, 2022; Xinogalos, 2021; Caspersen; Bennedsen, 2007), programação impõe alta *ICL* (Qian; Lehman, 2017), enquanto má instrução aumenta *ECL* (Mayer; Moreno, 2003). Estratégias como Exemplos Resolvidos (*WEs)* (Sweller; Chandler, 1994) e retirada-grau-a-grau-do-suporte ajudam a gerenciar carga. Efeito de expertise reversa (Kalyuga *et al.*, 2003) destaca necessidade de adaptação (considerando estudos de Chase e Simon, 1973). Tópicos de vanguarda em *AI* (ajuste fino, de *LMASs*) elevam *ICL* e *ECL* devido à complexidade conceitual/procedural e ferramentas imaturas, exigindo *design* instrucional para otimizar *GCL* via metacognição e prática, apoiado por neurociência educacional (Gkintoni *et al.*, 2025).

Ensino de ajuste fino local e orquestração de *LMASs* (Luo *et al.*, 2025) apresenta desafios pedagógicos, amplificando *CL* e introduzindo *CompL* como fator interdependente. Ajuste fino local exige compreensão de conceitos complexos (transferência de aprendizado, quantização - *INT8*/*INT4* discutidos por Nagel *et al.*, 2021; Yao *et al.*, 2022; Kim *et al.*, 2023; Zhao *et al.*, 2025; *QLoRA*, Dettmers *et al.*, 2023), elevando *ICL*. Este tipo de prática envolve tarefas cognitivamente exigentes (curadoria, hiperparâmetros, interpretação de perdas instáveis, especialmente com Q4), cuja instabilidade aumenta *ICL* e *ECL* (ligado a desafios de raciocínio de modelos - Chen *et al.*, 2025; Pu *et al.*, 2025; Kumar *et al.*, 2025; Wang *et al.*, 2025a; Xia *et al.*, 2025). Interação com *CompL* é crítica em *hardware* limitado (Bai *et al.*, 2024; Husom *et*

al., 2025; Jia *et al.*, 2025; Ni *et al.*, 2025), forçando balanços (Q4 *vs.* Q8 - H1) entre uso de *VRAM* e estabilidade/*CL*. Orquestração de *LMASs* herda complexidades de sistemas distribuídos (Andrews, 2000; Ben-Ari, 2006), elevando *ICL*, agravada por natureza probabilística de *LLMs* (exigindo *design* de *prompts* - Spiess *et al.*, 2025; e protocolos - Ghosh, 2025; Ahmadi *et al.*, 2024). Depuração complexa (*LangChain*, etc. - Wu *et al.*, 2023; Zhang *et al.*, 2025; OpenAI, 2025; Luo *et al.*, 2025; Truong *et al.*, 2023) aumentam *ECL*, enquanto execução paralela eleva *CompL* agregada. Literatura carece de abordagens sistemáticas para ensinar esses tópicos considerando *CL-CompL* e avaliação (Liu *et al.*, 2025b; Kim *et al.*, 2025).

Interdependência *CL-CompL* não foi ainda adequadamente teorizada na pesquisa educacional. Achados empíricos (H1-H4) indicam otimizações de *CompL* (Q4) aumentam *CL*, e tarefas complexas (*LMASs*) elevam ambas. Relação não linear sugere "ponto ideal" (H3) onde *CompL*/Qualidade moderada minimiza *CL*. Iniciantes (H4) são mais vulneráveis à sobrecarga, especialmente com restrições de *CompL*. Ignorar interação leva a sobrecarga cognitiva ou barreiras computacionais. Literatura raramente aborda interação cognição-recursos (Paas *et al.*, 2003; Zhang; Norman, 1994) no *design* instrucional, apesar da importância da interação humano-AI (Amershi *et al.*, 2019; Heer, 2019; Hoc, 2000; Xu, 2019; Retzlaff *et al.*, 2024).

A *AI* transforma a *SE* (Hassan *et al.*, 2024), mas políticas educacionais (PNE - Brasil, 2014; DCNs - Brasil, 2016, 2019) e pesquisa pedagógica ainda não acompanham, especialmente para tópicos emergentes. *CLT* (Sweller *et al.*, 1998; Duran *et al.*, 2022) é útil, mas desafios de ajuste fino (Liu *et al.*, 2023) e de *LMASs* (Luo *et al.*, 2025) introduzem a intrincada complexidade da interação *CL-CompL* em ambientes com recursos limitados (Bai *et al.*, 2024) como lacuna crítica. Este trabalho propõe preencher esta lacuna com um modelo pedagógico que gerencie balanço CL-CompL (H1), sobrecarga de orquestração (H2), busca pelo ponto ideal entre cognição e computabilidade (H3) e adaptação à expertise (H4, Kalyuga *et al.*, 2003), visando a formação em *SE* mais eficaz e equitativa na era da *AI*.

ANÁLISE DE DADOS

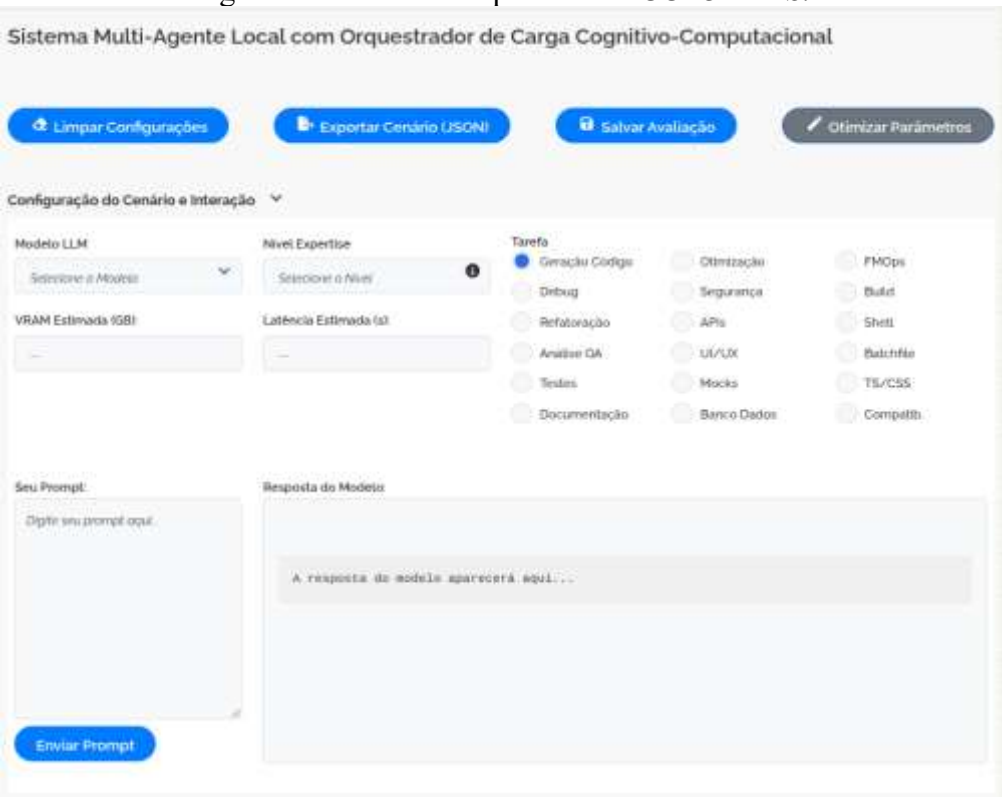
Detalha-se o processo analítico interno para avaliar, refinar e validar conceitualmente o modelo pedagógico proposto para integrar tópicos de ponta em *AI* (*LMAS*, *QLoRA*) no ensino de SE. Este estudo focou na reflexão crítica e sistemática da proposta, sem novos dados, baseando-se em teorias educacionais (*CLT* - Sweller *et al.*, 1998; *HCI* - Nielsen, 1993; Norman, 1988), achados empíricos prévios sobre a dinâmica *CL-CompL* (hipóteses H1-H4) e

considerações contextuais (limitações de *hardware* no Brasil). O objetivo foi assegurar um modelo teoricamente sólido, coerente, empiricamente sustentado e viável para implementação.

O processo analítico teve três fases: definição de critérios para avaliar a proposta (coerência teórica, fundamentação empírica, viabilidade prática, impacto pedagógico); aplicação sistemática desses critérios a cada componente da proposta – princípios, estratégias curriculares, atividades práticas (fundamentadas por pedagogias ativas como as de Hmelo-Silver, 2004 e Thomas, 2000), ferramentas de apoio, avaliação e adaptação; e mapeamento conectando cada estratégia aos achados empíricos (H1-H4), fundamentos teóricos (*CLT*, *HCI*, pedagogia ativa) e objetivos de aprendizagem, garantindo rastreabilidade e fundamentos.

A análise enfatizou a gestão da *CompL*, sugerindo soluções como uso de modelos de linguagem moderados, quantização estratégica (Q4/Q8) com suporte pedagógico, e execução assíncrona de tarefas computacionalmente intensivas, transformando limitações em oportunidades de aprendizado. A estrutura final do modelo pedagógico foi organizada em categorias funcionais – princípios gerais, estratégias curriculares, *design* de atividades, recomendações para ferramentas (plataforma *CCLOLMAS*, figura abaixo), abordagens de avaliação e adaptação à expertise – facilitando sua adoção.

Figura 1 – Interface da plataforma *CCLOLMAS*.



Fonte: Plataforma *CCLOLMAS* (2025).

O processo consolidou os insights em proposta pedagógica estruturada, coerente e pragmática, capaz de equilibrar desafios teóricos, empíricos e contextuais. A ênfase na interação *CL-CompL*, combinada com estratégias de ensino ativas e contextualizadas, visa preparar estudantes para enfrentar balanços reais na aplicação de *AI* em SE, fomentando aprendizado e competências críticas.

RESULTADOS

Apresenta-se o modelo pedagógico para integrar tópicos de *AI* emergentes (*LMAS*; *QLoRA*, Dettmers *et al.*, 2023) em currículos de SE. Estes resultados emergem da síntese entre achados empíricos sobre a interdependência *CL-CompL*, a *CLT* (Sweller *et al.*, 1998), princípios de *HCI* (Nielsen, 1993; Norman, 1988; Amershi *et al.*, 2019) e metodologias ativas (Hmelo-Silver, 2004; Thomas, 2000). O objetivo é preencher lacunas oferecendo uma abordagem que gerencie a dinâmica *CL-CompL* em ambientes com recursos limitados. O modelo se estrutura em cinco princípios: (1) Reconhecimento da interdependência *CL-CompL* (baseado em H1, H2, H3), onde decisões técnicas (Q4 *vs.* Q8) impactam *CL*, exigindo atividades reflexivas e ferramentas com *feedback* duplo (*VRAM*, esforço percebido - *SEQ*/Paas *et al.*, 2003). (2) Foco no "ponto ideal" como objetivo metacognitivo (derivado de H3), desenvolvendo capacidade de avaliar balanços e justificar escolhas (ex.: estrutura de *LMASs*, quantização). (3) Minimização de *ECL* e otimização de *GCL* via *design* instrucional e usabilidade (Mayer e Moreno, 2003), com interfaces intuitivas, *WEs* e *feedback* formativo (Shute, 2008). (4) Adaptação à expertise e restrições de *hardware* (evitando o efeito de reversão da expertise), com sequenciamento gradual, estrutura de apoio diferenciada (Wood *et al.*, 1976) e alternativas para baixa *CompL* (ex.: Phi-4-mini). (5) Integração da *AI* a práticas autênticas de *SE* via *PBL* (Hmelo-Silver, 2004) ou *PjBL* (Thomas, 2000), com projetos como ajuste fino para tradução de código ou agentes para revisão (Afrin *et al.*, 2025), contextualizando a gestão *CL-CompL* dentro de práticas emergentes como *LLMOps* (Pahune e Akhtar, 2025; Tantithamthavorn *et al.*, 2025).

O Sequenciamento Curricular propõe progressão gradual: engenharia de *prompt* -> quantização -> execução local -> ajuste fino (Ayupov; Chirkova, 2022; Liu *et al.*, 2023) -> *QLoRA*/orquestração de *LMASs*, com complexidade crescente de ferramentas (LM Studio -> bibliotecas). O *Design* de Atividades Práticas inclui laboratórios comparativos (Q4 *vs.* Q8), iniciando com Q8 para estabilidade e introduzindo Q4 com estrutura de apoio; para *LMASs*, começando com fluxos simples usando *frameworks* (*CrewAI* - Luo *et al.*, 2025) e focando em

depuração; projetos autênticos (geração de testes) contextualizam a aplicação. Ferramentas de apoio, como a plataforma *CCLOLMAS* (*React*/*FastAPI*) com monitoramento de *CompL*, visualizações e/ou uso guiado por outras ferramentas de código aberto (*Ollama* - Wu *et al.*, 2023) adaptadas a restrições. Avaliação combina métodos formativos (observação, logs, reflexão - Shute, 2008) e somativos (qualidade, depuração), focando no processo e ciente dos desafios de avaliação na era da *AI* (Tate *et al.*, 2023). Adaptação à Expertise e Contexto envolve diagnóstico inicial, flexibilidade de ferramentas, adaptação a *hardware* (modelos menores) e gestão de expectativas via execução assíncrona.

A síntese destes resultados ressalta a inovação ao centralizar a gestão da interdependência *CL-CompL* no *design* instrucional. A proposta capacita educadores a lidar com os balanços técnicos (eficiência *vs.* qualidade), mitigar sobrecarga cognitiva em tarefas complexas (*QLoRA*, *LMASs*) e desenvolver metacognição para identificar o "ponto ideal". A adaptação a diversas realidades (expertise, *hardware*) é viabilizada por sequenciamento, por uma estrutura de apoio (Wood *et al.*, 1976) e execução assíncrona. Integração com práticas autênticas de *SE* via *PjBL* (Thomas, 2000) torna o aprendizado significativo e transferível para desafios industriais (automação de testes). A implementação visa formar engenheiros de *software* mais preparados para a era da *AI*, capazes de gerenciar balanços, adaptar-se e aplicar *AI* de forma sustentável, com implicações discutidas adiante.

DISCUSSÃO

Discute-se a urgência de uma transformação pedagógica no ensino de Engenharia de *software* (*SE*), centrada na interação dinâmica entre Carga Cognitiva (*CL*) e Carga Computacional (*CompL*), para integrar eficazmente tópicos sofisticados em Inteligência Artificial (*AI*), como a orquestração de Sistemas Multiagentes Locais (*LMASs*) (Luo *et al.*, 2025; Wu *et al.*, 2023; Zhang *et al.*, 2025) e ajuste fino local (ex.: *QLoRA*, Dettmers *et al.*, 2023). Essa abordagem, fundamentada em achados empíricos (H1-H4), propõe um modelo conceitual que supera a visão tradicional de ensino focado apenas no "o quê" e no "como usar" ferramentas de *AI*, integrando a gestão explícita dos balanços entre as limitações cognitivas do aprendiz e as restrições computacionais do ambiente.

A Teoria da Carga Cognitiva (*CLT*) (Sweller *et al.*, 1998; Van Merrienboer; Sweller, 2005), amplamente aplicada em computação (Duran *et al.*, 2022; Xinogalos, 2021), é recontextualizada ao reconhecer que a *CompL* — como uso de *VRAM*, tempo de execução e consumo energético — não é um mero obstáculo técnico, mas um fator interativo que modula

diretamente a *CL*. Por exemplo, a otimização excessiva da *CompL* via quantização agressiva (Q4) (Nagel *et al.*, 2021) pode reduzir recursos exigidos, mas gera instabilidade e baixa qualidade de resultados, aumentando a *CL* devido à necessidade de depuração e validação contínuas. Inversamente, ignorar limitações computacionais pode tornar atividades práticas inviáveis, quebrando o fluxo de aprendizado. O "ponto ideal" entre essas cargas é apresentado como um empreendimento cognitivo essencial, exigindo que estudantes analisem dados, justifiquem decisões e equilibrem balanços como desempenho técnico, custo computacional e esforço cognitivo — habilidades críticas para engenheiros que atuarão em sistemas híbridos humano-AI (Amershi *et al.*, 2019; Heer, 2019; Hoc, 2000; Xu, 2019).

Essa perspectiva implica mudanças profundas no currículo de SE. Em vez de adicionar disciplinas isoladas de *AI*, sugere-se integrar a gestão *CL-CompL* transversalmente, como em Arquitetura de *software*, onde a escolha de modelos de *AI* deve considerar não apenas requisitos técnicos (latência, *hardware*) mas também o impacto na complexidade de integração e manutenção (considerando práticas de *LLMOps*, Pahune e Akhtar, 2025; Tantithamthavorn *et al.*, 2025), afetando a *CL* da equipe.

Em Gerência de Projetos, é necessário estimar recursos computacionais para tarefas de *AI* e planejar a gestão do esforço cognitivo diante de ferramentas instáveis ou em constante evolução. Em Interação Humano-Computador (*HCI*), discute-se o *design* de interfaces que minimizem a frustração ao lidar com assistentes de código ou plataformas MLOps (Retzlaff *et al.*, 2024). Novos conteúdos, como orquestração de agentes (ex.: padrão *Master-Worker*), gerenciamento de recursos (monitoramento de *VRAM*, estratégias de quantização) e metacognição sobre balanços, devem ser incorporados. Laboratórios práticos exigem reconfiguração para explorar esses balanços em um ambiente austero de *hardware* limitado, usando modelos menores (*Phi-4-mini*, *Mistral* e *DeepSeek-R1* - Jiang *et al.*, 2023; Marjanovic *et al.*, 2025) e quantização criteriosa, com ênfase na análise crítica do processo, não apenas no resultado final. A flexibilidade curricular permitida pelas Diretrizes Curriculares Nacionais (DCNs) é um recurso importante, mas sua aplicação requer planejamento estratégico para evitar abordagens superficiais.

A implementação dessa pedagogia depende da formação docente. Professores precisam dominar tecnologias emergentes (*QLoRA*, *LMASs*), compreender a *CLT* e suas implicações no *design* instrucional, e enriquecer seu repertório didático com estratégias para gerenciar a interação *CL-CompL*. Programas de desenvolvimento profissional contínuo devem abordar atualização tecnológica, aprofundamento em *CLT*, competências em gerenciamento da *CompL*

e uso de ferramentas educacionais dedicadas, como a plataforma *CCLOLMAS* aqui proposta. A produção de materiais de apoio (planos de aula testados, modelos pré-quantizados) é essencial para facilitar a adoção em larga escala. Superar resistências exige apresentar o modelo *CL-CompL* não como uma complicação adicional, mas como uma ferramenta para estruturar a complexidade da *AI*, tornando o ensino mais eficaz e menos intimidador.

O desenvolvimento de ferramentas educacionais também é central. Plataformas dedicadas devem abstrair complexidade desnecessária (ex.: interfaces intuitivas para ajuste fino), visualizar métricas de *CL-CompL*, gerenciar recursos de forma transparente (alertas de configurações problemáticas) e integrar uma estrutura de apoio pedagógico (Wood *et al.*, 1976) (exemplos comentados, *feedback just-in-time*). A adaptação de ferramentas de código aberto, como *Ollama*, requer tutoriais passo a passo, notebooks pré-configurados e scripts para monitoramento de *CompL*, reduzindo a Carga Extrínseca (*ECL*).

A curadoria de modelos (ex.: Q4/Q8, 7B-35B parâmetros) e *datasets* relevantes para *SE* (geração de código, sumarização - Afrin *et al.*, 2025) é primordial para atividades práticas que permitam explorar os balanços *CL-CompL* de forma significativa, mesmo em equipamentos com recursos limitados. Modelos e *datasets* devem ser testados e validados quanto à acessibilidade, relevância para tarefas de engenharia de *software* (como geração de código, classificação de mensagens de commit ou sumarização de código) e capacidade de ilustrar os impactos de diferentes configurações de quantização ou arquiteturas de agentes.

As implicações para políticas educacionais, como o Plano Nacional de Educação (PNE) (Brasil, 2014; Brasil, 2023; Brasil, 2024) e as Diretrizes Curriculares Nacionais (DCNs) (Brasil, 2016; Brasil, 2019), são igualmente profundas. Além de reconhecer a dinâmica Cognição Híbrida-Computação Híbrida (*CL-CompL*) como uma competência transversal essencial, as políticas devem promover uma formação metacognitiva crítica, que vá além do uso instrumental de ferramentas de *AI*. Isso se alinha diretamente às "Diretrizes comuns da Avaliação de Permanência dos Programas de Pós-Graduação *stricto sensu"* (Brasil, 2025), que enfatizam, na sua análise multidimensional e através da Ficha de Avaliação (especialmente nos Quesitos 1 - Programa e 2 - Formação), a necessidade de "boas práticas de formação, pesquisa e disseminação científica" e a "integridade científica". Ensinar estudantes a avaliar custos cognitivos/computacionais e refletir sobre a interação humano-IA (Floridi; Taddeo, 2016; Jobin *et al.*, 2019) e estratégicas torna-se estruturante, impactando a avaliação da "Qualidade das teses, dissertações ou equivalentes" (Item 2.1) e das "atividades de pesquisa" (Item 2.4). A infraestrutura computacional, conforme abordado no Item 1.1 ("Identidade e condições de

funcionamento do Programa: corpo docente, infraestrutura, estrutura curricular, etc"), deve ser tratada como questão pedagógica central, garantindo equidade no acesso a recursos como *GPUs*, o que dialoga com a dimensão de "Ações afirmativas, de inclusão, permanência e acessibilidade" das Diretrizes. Políticas de formação docente (PNE Meta 16) devem capacitar professores para a gestão pedagógica da *CL-CompL*, integrando essas ações ao "Planejamento estratégico do Programa" (Item 1.3) e ao PDI institucional, visando não apenas a formação qualificada, mas também o "Impacto" (Quesito 3) e a "Interação com a sociedade", aspectos valorizados pela avaliação de Brasil (2025).

A capacidade de trabalhar com *AI* eficazmente — compreendendo seus mecanismos, avaliando suas limitações e gerenciando os complexos balanços entre custos cognitivos e computacionais — será uma habilidade definidora fundamental para profissionais nas próximas décadas (Hassan *et al.*, 2024). A *SE* está se tornando uma disciplina de colaboração cada vez mais simbiótica entre humanos e sistemas de *AI* (Hoc, 2000; Heer, 2019), onde a capacidade de lidar efetivamente a complexidade da interação *CL-CompL* determinará a eficácia, a inovação e a sustentabilidade das soluções desenvolvidas. Uma pedagogia que ignore essa dinâmica corre o risco de formar profissionais meramente reativos a ferramentas de *AI*, incapazes de lidar com a instabilidade, a complexidade ou os custos associados. Em contraste, a abordagem proposta neste trabalho — que ensina estudantes a buscar o "ponto ideal" entre *CL* e *CompL*, a adaptar estratégias conforme a expertise (Kalyuga *et al.*, 2003) e as restrições do ambiente, assim como a refletir criticamente sobre suas escolhas — tem o potencial de formar engenheiros mais resilientes e adaptáveis.

Em síntese, reforça-se a necessidade urgente de uma mudança de paradigma no ensino de *AI* dentro dos cursos de SE. A dinâmica *CL-CompL* não é um detalhe secundário, mas um fenômeno central que deve sustentar o *design* curricular, as práticas de ensino, o desenvolvimento de ferramentas educacionais e as políticas educacionais. Superar os desafios de implementação — desde a atualização curricular e a formação docente até o investimento em infraestrutura e o desenvolvimento de recursos educacionais — é essencial para preparar estudantes para a realidade inescapável da engenharia de *software* na era da inteligência artificial. A integração consciente e sistemática dessa perspectiva pedagógica não apenas melhora a qualidade do ensino, mas também alinha a formação profissional às demandas do mercado e às exigências sociais de uma tecnologia que redefine continuamente o engenho humano.

CONSIDERAÇÕES FINAIS

Este estudo enfrentou o desafio de integrar avanços disruptivos da *AI*, como orquestração de *LMASs* (Luo *et al.*, 2025) e técnicas de quantização (*QLoRA*, Dettmers *et al.*, 2023; Nagel *et al.*, 2021), em currículos densos de Engenharia de *software* (*SE*), especialmente com recursos limitados e falta de diretrizes políticas no Brasil. Propõe-se um modelo teórico-prático focado na gestão ativa da interdependência entre Carga Cognitiva (*CL*) do estudante (Sweller *et al.*, 1998) e Carga Computacional (*CompL*) do ambiente.

A investigação empírica revelou que otimizar apenas a *CompL* (ex.: quantização Q4) eleva a *CL* devido à instabilidade (H1), e a complexidade dos *LMASs* amplifica ambas as cargas (H2). A relação não linear sugere um "ponto ideal" (H3), onde configurações mais estáveis (Q8) reduzem a *CL* mesmo com maior *CompL*. A expertise modera o efeito (H4), com iniciantes mais suscetíveis à sobrecarga (Kalyuga *et al.*, 2003). A contribuição central é adaptar esses achados à pedagogia, tornando a dinâmica *CL-CompL* um núcleo estruturante do ensino de *AI* em *SE* e promovendo metacognição para avaliar balanços.

O modelo proposto baseia-se em cinco princípios: reconhecimento explícito da interdependência *CL-CompL*, desenvolvimento de habilidades para o equilíbrio ideal, *design* instrucional via *CLT* (minimizar carga extrínseca, otimizar germânica), adaptação flexível (expertise/*hardware*) e integração com práticas autênticas de *SE* (metodologias ativas). Estratégias incluem sequenciamento progressivo, atividades comparativas (Q4/Q8, de *LMASs*), ferramentas para visualizar a dinâmica *CL-CompL* e avaliação focada no processo reflexivo. O objetivo é formar profissionais que reconheçam os custos cognitivos/computacionais da *AI*.

Trabalhos futuros devem validar empiricamente as estratégias propostas e explorar novas e promissoras aplicações. Em suma, este estudo argumenta que a nevrálgica sinergia mente-máquina, gerenciando ativamente a complexa interação *CL-CompL*, é essencial para capacitar engenheiros de *software* a liderar inovações éticas (Floridi; Taddeo, 2016) e responsáveis na atual era da *AI*, sendo um investimento estratégico vital para o Brasil.

REFERÊNCIAS

AFRIN, S. *et al*. **Resource-Efficient & Effective Code Summarization**. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2502.03617. Acesso em: 20 abr. 2025.

AHMADI, A.; SHARIF, S.; BANAD, Y. M. **MCP Bridge**: A Lightweight, LLM-Agnostic RESTful Proxy for Model Context Protocol Servers. [S. l.]: arXiv, 2024. Disponível em: https://arxiv.org/abs/2504.08999. Acesso em: 20 abr. 2025.

AMERSHI, S. *et al*. Guidelines for Human-AI Interaction. *In*: CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI '19), 2019, Glasgow. **Proceedings [...]**. New York: ACM, 2019. p. 1–13. DOI: 10.1145/3290605.3300233. Disponível em: https://doi.org/10.1145/3290605.3300233. Acesso em: 20 abr. 2025.

ANDREWS, G. R. **Foundations of Multithreaded, Parallel, and Distributed Programming**. Reading, MA: Addison-Wesley, 2000.

AYUPOV, S.; CHIRKOVA, N. **Parameter-efficient finetuning of transformers for source code**. [S. l.]: arXiv, 2022. Disponível em: https://arxiv.org/abs/2212.05901. Acesso em: 20 abr. 2025.

BAI, G. *et al*. **Beyond Efficiency**: A Systematic Survey of Resource-Efficient Large Language Models. [S. l.]: arXiv, 2024. Disponível em: https://arxiv.org/abs/2401.00625. Acesso em: 20 abr. 2025.

BEN-ARI, M. **Principles of Concurrent and Distributed Programming**. 2. ed. Harlow: Addison-Wesley, 2006.

BIRD, S. *et al*. Taking Flight with Copilot: Early Insights and Opportunities of Large Language Models in Programming Education. *In*: ACM TECHNICAL SYMPOSIUM ON COMPUTER SCIENCE EDUCATION (SIGCSE '22), 2022, Providence. **Proceedings [...]**. New York: ACM, 2022. p. 1032–1032. DOI: 10.1145/3478431.3499901.

BIRKMOSE, R. *et al*. **On-Device LLMs for Home Assistant**: Dual Role in Intent Detection and Response Generation. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2502.12923. Acesso em: 20 abr. 2025.

BRASIL. **Lei nº 13.005, de 25 de junho de 2014**. Aprova o Plano Nacional de Educação - PNE e dá outras providências. Diário Oficial da União: seção 1, Brasília, DF, ed. extra, p. 1, 26 jun. 2014. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l13005.htm. Acesso em: 25 abr. 2025.

BRASIL. [Ministério da Educação]. **Países dos Brics discutem avaliação da educação superior**. Brasília, DF: MEC, 2025. Disponível em: https://www.gov.br/mec/pt-br/assuntos/noticias/2025/abril/paises-dos-brics-discutem-avaliacao-da-educacao-superior. Acesso em: 3 maio 2025.

BRASIL. Conselho Nacional de Educação. Câmara de Educação Superior. **Resolução CNE/CES nº 5, de 16 de novembro de 2016**. Institui as Diretrizes Curriculares Nacionais para os cursos de graduação na área da Computação [...]. Brasília, DF: MEC, 2016. Disponível em: http://portal.mec.gov.br/docman/novembro-2016-pdf/52101-rces005-16-pdf/file. Acesso em: 25 abr. 2025.

BRASIL. Conselho Nacional de Educação. Câmara de Educação Superior. **Resolução CNE/CES nº 2, de 24 de abril de 2019**. Institui as Diretrizes Curriculares Nacionais do Curso de Graduação em Engenharia. Diário Oficial da União: seção 1, Brasília, DF, p. 42, 26 abr. 2019. Disponível em: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=112091-rces002-19&category_slug=abril-2019-pdf&Itemid=30192. Acesso em: 25 abr. 2025.

BRASIL. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. **Diretrizes comuns da Avaliação de Permanência dos Programas de Pós-Graduação stricto sensu**: Ciclo avaliativo 2025-2028, Avaliação Quadrienal 2029. Brasília, DF: CAPES, 2025. DOI 10.21713/Diretrizescomuns. Disponível em: https://uploads.capes.gov.br/files/2025-05-02_DocumentoReferencial_FICHA.pdf. Acesso em: 3 maio 2025.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. **Relatório do 2º Ciclo de Monitoramento das Metas do Plano Nacional de Educação 2022**. Brasília, DF: INEP, 2023. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/pne/monitoramento-continuo. Acesso em: 25 abr. 2025.

BRASIL. Ministério da Educação. **Conferência Nacional de Educação (CONAE) 2024**. Brasília, DF: MEC, 2024. Disponível em: https://www.gov.br/mec/pt-br/acesso-a-informacao/participacao-social/conferencias/conae-2024. Acesso em: 20 abr. 2025.

CASPERSEN, M. E.; BENNEDSEN, J. Instructional design of programming education: a learning theoretic approach. **ACM SIGCSE Bulletin**, v. 39, n. 4, p. 111-124, Dec. 2007. DOI: 10.1145/1345375.1345413.

CHASE, W. G.; SIMON, H. A. Perception in chess. **Cognitive Psychology**, v. 4, n. 1, p. 55-81, 1973. DOI: 10.1016/0010-0285(73)90004-2. Disponível em: https://doi.org/10.1016/0010-0285(73)90004-2. Acesso em: 20 abr. 2025.

CHEN, D. *et al*. **Overthinking Leads to Errors**: Mitigating Overthinking in Large Language Models through Task Difficulty Calibration. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2504.13367. Acesso em: 24 abr. 2025.

DENNING, P. J.; TEDRE, M. **Computational Thinking**. Cambridge, MA: MIT Press, 2019.

DETTMERS, T. *et al*. Qlora: Efficient finetuning of quantized llms. **Advances in Neural Information Processing Systems**, v. 36, 2023. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2023/hash/6a4d5956f508541689d0e5c5a878e074-Abstract-Conference.html. Acesso em: 20 abr. 2025.

DURAN, R.; ZAVGORODNIAIA, A.; SORVA, J. Cognitive Load Theory in Computing Education Research: A Review. **ACM Transactions on Computing Education**, v. 22, n. 4, art. 40, p. 1–27, 2022. DOI: 10.1145/3532776. Disponível em: https://doi.org/10.1145/3532776. Acesso em: 20 abr. 2025.

ENGESTROM, Y. **Learning by expanding**: An activity-theoretical approach to developmental research. Helsinki: Orienta-Konsultit, 1987.

FERRAG, Mohamed Amine; TIHANYI, Norbert; DEBBAH, Merouane. **From LLM Reasoning to Autonomous AI Agents**: A Comprehensive Review. [S. l.]: arXiv, 28 abr. 2025. Disponível em: https://arxiv.org/pdf/2504.19678. Acesso em: 5 maio 2025.

FLORIDI, L.; TADDEO, M. What is data ethics? **Philosophical Transactions of the Royal Society A**: Mathematical, Physical and Engineering Sciences, v. 374, n. 2083, p. 20160360,

2016. DOI: 10.1098/rsta.2016.0360. Disponível em: https://doi.org/10.1098/rsta.2016.0360. Acesso em: 20 abr. 2025.

GAO, Xin *et al*. **A Strategic Coordination Framework of Small LLMs Matches Large LLMs in Data Synthesis**. [S. l.]: arXiv, 21 abr. 2025. Disponível em: https://arxiv.org/html/2504.12322v2. Acesso em: 5 maio 2025.

GHOSH, D. P. **Agentic ecosystem in engineering design**: a framework for interoperable legacy tools and emergent collaboration via MCP/A2A protocols. Kolkata: Engineering Design & Research Center, Larsen & Toubro Construction, 2025. Disponível em: https://www.researchgate.net/publication/390953589. Acesso em: 20 abr. 2025.

GHOSH, Dibya; GHOSH, Dyuti; GHOSH, Debi Prasad. **Think in Arrows**: A Categorical Scaffolding Framework for Robust Artificial Scientific Discovery. [S. l.]: ResearchGate, abr. 2025. Preprint. DOI: 10.13140/RG.2.2.16950.41280. Acesso em: 5 maio 2025.

GKINTONI, E. *et al*. Challenging Cognitive Load Theory: The Role of Educational Neuroscience and Artificial Intelligence in Redefining Learning Efficacy. **Brain Sciences**, v. 15, n. 2, p. 203, 2025. DOI: 10.3390/brainsci15020203. Disponível em: https://doi.org/10.3390/brainsci15020203. Acesso em: 20 abr. 2025.

HASSAN, Ahmed E. *et al*. Rethinking Software Engineering in the Era of Foundation Models: A Curated Catalogue of Challenges in the Development of Trustworthy FMware. *In*: INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING COMPANION (ICSE '24 Companion), 2024, Lisbon. **Proceedings [...]**. New York: ACM, 2024. p. 1010-1019. DOI: 10.1145/3663529.3663849.

HEER, J. Agency plus automation: Designing artificial intelligence into interactive systems. **Proceedings of the National Academy of Sciences**, v. 116, n. 6, p. 1844-1850, 2019. DOI: 10.1073/pnas.1807180115. Disponível em: https://doi.org/10.1073/pnas.1807180115. Acesso em: 20 abr. 2025.

HMELO-SILVER, C. E. Problem-Based Learning: What and How Do Students Learn? **Educational Psychology Review**, v. 16, n. 3, p. 235-266, 2004. DOI: 10.1023/B:EDPR.0000034022.16470.f3.

HOC, J.-M. From human-machine interaction to human-machine cooperation. **Ergonomics**, v. 43, n. 7, p. 833-843, 2000. DOI: 10.1080/001401300409060. Disponível em: https://doi.org/10.1080/001401300409060. Acesso em: 20 abr. 2025.

HUSOM, E. J. *et al*. **Sustainable LLM Inference for Edge AI**: Evaluating Quantized LLMs for Energy Efficiency, Output Accuracy, and Inference Latency. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2504.03360. Acesso em: 20 abr. 2025.

HUTCHINS, E. **Cognition in the Wild**. Cambridge, MA: MIT Press, 1995.

JIA, F. *et al*. **Scaling Up On-Device LLMs via Active-Weight Swapping Between DRAM and Flash**. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2504.08378. Acesso em: 20 abr. 2025.

JIANG, A. Q. *et al*. **Mistral 7B**. [S. l.]: arXiv, 2023. Disponível em: https://arxiv.org/abs/2310.06825. Acesso em: 20 abr. 2025.

JOBIN, A.; IENCA, M.; VAYENA, E. The global landscape of AI ethics guidelines. **Nature Machine Intelligence**, v. 1, n. 9, p. 389-399, 2019. DOI: 10.1038/s42256-019-0088-2. Disponível em: https://doi.org/10.1038/s42256-019-0088-2. Acesso em: 20 abr. 2025.

KALYUGA, S. *et al*. The expertise reversal effect. **Educational Psychologist**, v. 38, n. 1, p. 23-31, 2003. DOI: 10.1207/S15326985EP3801_4. Disponível em: https://doi.org/10.1207/S15326985EP3801_4. Acesso em: 20 abr. 2025.

KIM, S. *et al*. **SqueezeLLM**: Dense-and-Sparse Quantization. [S. l.]: arXiv, 2023. Disponível em: https://arxiv.org/abs/2306.07629. Acesso em: 20 abr. 2025.

KIM, Y. *et al*. **One ruler to measure them all**: Benchmarking multilingual long-context language models. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2503.01996. Acesso em: 20 abr. 2025.

KOTALWAR, N.; GOTOVOS, A.; SINGLA, A. **Hints-In-Browser**: Benchmarking Language Models for Programming Feedback Generation. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2406.05053. Acesso em: 20 abr. 2025.

KUMAR, K. *et al*. **LLM Post-Training**: A Deep Dive into Reasoning Large Language Models. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2502.21321. Acesso em: 20 abr. 2025.

LEINONEN, J. *et al*. Using Large Language Models to Enhance Programming Education. *In*: ACM CONFERENCE ON INTERNATIONAL COMPUTING EDUCATION RESEARCH (ICER '23), 2023, Chicago. **Proceedings [...]**. New York: ACM, 2023. p. 1–15. DOI: 10.1145/3568813.3600141.

LI, R. *et al*. **StarCoder**: may the source be with you! [S. l.]: arXiv, 2023. Disponível em: https://arxiv.org/abs/2305.06161. Acesso em: 25 abr. 2025.

LIN, Kevin *et al*. **Sleep-time Compute**: Beyond Inference Scaling at Test-time. [S. l.]: arXiv, 17 abr. 2025. Preprint. Disponível em: https://arxiv.org/html/2504.13171v1. Acesso em: 3 maio 2025.

LIU, J.; SHA, C.; PENG, X. An empirical study of parameter-efficient fine-tuning methods for pre-trained code models. *In*: IEEE/ACM INTERNATIONAL CONFERENCE ON AUTOMATED SOFTWARE ENGINEERING (ASE), 38., 2023, Kirchberg. **Proceedings [...]**. Piscataway: IEEE, 2023. p. 397–408. DOI: 10.1109/ASE56229.2023.00041. Acesso em: 20 abr. 2025.

LIU, Na *et al*. **From LLM to Conversational Agent**: A Memory Enhanced Architecture with Fine-Tuning of Large Language Models. [S. l.]: arXiv, 30 jan. 2024. Disponível em: https://arxiv.org/pdf/2401.02777. Acesso em: 5 abr. 2025.

LIU, Y. *et al*. **AgentBench**: Evaluating LLMs as Agents. [S. l.]: arXiv, 2025b. Disponível em: https://arxiv.org/abs/2308.03688. Acesso em: 20 abr. 2025.

LUAN, H. *et al*. **Who Needs to Learn Programming Nowadays?** The Impact of AI on Programming Education. [S. l.]: arXiv, 2023. Disponível em: https://arxiv.org/abs/2305.10881. Acesso em: 25 abr. 2025.

LUO, J. *et al*. **Large Language Model Agent**: A Survey on Methodology, Applications and Challenges. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2503.21460. Acesso em: 20 abr. 2025.

MA, Hao *et al*. **Coevolving with the Other You**: Fine-Tuning LLM with Sequential Cooperative Multi-Agent Reinforcement Learning. [S. l.]: arXiv, 22 fev. 2025. Disponível em: https://arxiv.org/html/2410.06101v2. Acesso em: 5 maio 2025.

MARJANOVIC, S. V. *et al*. **DeepSeek-R1 Thoughtology**: Let's about LLM reasoning. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2504.07128. Acesso em: 20 abr. 2025.

MAYER, R. E.; MORENO, R. Nine ways to reduce cognitive load in multimedia learning. **Educational Psychologist**, v. 38, n. 1, p. 43-52, 2003. DOI: 10.1207/S15326985EP3801_6. Disponível em: https://doi.org/10.1207/S15326985EP3801_6. Acesso em: 20 abr. 2025.

MHLANGA, D. The Role of Artificial Intelligence in Enhancing Education Quality in Developing Countries. **Sustainability**, v. 15, n. 15, p. 11608, 2023. DOI: 10.3390/su151511608.

MISHRA, P.; KOEHLER, M. J. Technological Pedagogical Content Knowledge: A framework for teacher knowledge. **Teachers College Record**, v. 108, n. 6, p. 1017–1054, Jun. 2006. DOI: 10.1111/j.1467-9620.2006.00684.x. Acesso em: 20 abr. 2025.

NAGEL, M. *et al*. **A White Paper on Neural Network Quantization**. [S. l.]: arXiv, 2021. Disponível em: https://arxiv.org/abs/2106.08295. Acesso em: 20 abr. 2025.

NI, J. *et al*. **From Large to Super-Tiny**: End-to-End Optimization for Cost-Efficient LLMs. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2504.13471. Acesso em: 24 abr. 2025.

NIELSEN, J. **Usability Engineering**. Boston: Academic Press, 1993.

NORMAN, D. A. **The psychology of everyday things**. New York: Basic Books, 1988.

OPENAI. **A practical guide to building agents**. OpenAI Business Guides and Resources, 2025. Disponível em: https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf. Acesso em: 20 abr. 2025.

PAAS, F. *et al*. Cognitive load measurement as a means to advance cognitive load theory. **Educational Psychologist**, v. 38, n. 1, p. 63-71, 2003. DOI: 10.1207/S15326985EP3801_8. Disponível em: https://doi.org/10.1207/S15326985EP3801_8. Acesso em: 20 abr. 2025.

PAHUNE, S.; AKHTAR, Z. Transitioning from MLOps to LLMOps: Navigating the Unique Challenges of Large Language Models. **Information**, v. 16, n. 2, 87, 2025. DOI: 10.3390/info16020087. Acesso em: 24 abr. 2025.

PERRY, N. *et al*. **Do Users Write More Insecure Code with AI Assistants?** [S. l.]: arXiv, 2022. Disponível em: https://arxiv.org/abs/2211.03622. Acesso em: 25 abr. 2025.

PICCOLO, F. *et al*. Exploring the Impact of Large Language Models on Novice Programming. *In*: ACM TECHNICAL SYMPOSIUM ON COMPUTER SCIENCE EDUCATION (SIGCSE '24), 2024, Portland. **Proceedings [...]**. New York: ACM, 2024. p. 100–106. DOI: 10.1145/3626252.3630901.

PU, X. *et al*. **ThoughtTerminator**: Benchmarking, Calibrating, and Mitigating Overthinking in Reasoning Models. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2504.13367. Acesso em: 24 abr. 2025.

QIAN, Y.; LEHMAN, J. Students' Misconceptions and Other Difficulties in Introductory Programming: A Literature Review. **ACM Transactions on Computing Education (TOCE)**, v. 18, n. 1, p. 1-24, 2017. DOI: 10.1145/3077608.

RAY, Partha Pratim. **A Survey on Model Context Protocol**: Architecture, State-of-the-art, Challenges and Future Directions. [S. l.]: TechRxiv, 2025. Preprint. DOI: 10.36227/techrxiv.174495492.22752319. Acesso em: 5 maio 2025.

RETZLAFF, M. *et al*. **Evaluating Large Language Models for HCI**. [S. l.]: arXiv, 2024. Disponível em: https://arxiv.org/abs/2402.01811. Acesso em: 20 abr. 2025.

ROZIERE, B. *et al*. **Code Llama**: Open Foundation Models for Code. [S. l.]: arXiv, 2023. Disponível em: https://arxiv.org/abs/2308.12950. Acesso em: 25 abr. 2025.

SHUTE, V. J. Focus on Formative Feedback. **Review of Educational Research**, v. 78, n. 1, p. 153-189, 2008. DOI: 10.3102/0034654307313795.

SIDDIQ, M. A. *et al*. **Generative AI in Software Engineering Education**: A Systematic Literature Review. [S. l.]: arXiv, 2024. Disponível em: https://arxiv.org/abs/2401.10841. Acesso em: 25 abr. 2025.

SPIESS, C. *et al*. **AutoPDL**: Automatic Prompt Optimization for LLM Agents. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2504.04365. Acesso em: 24 abr. 2025.

SWELLER, J. Cognitive load theory, learning difficulty, and instructional design. **Learning and Instruction**, v. 4, n. 4, p. 295-312, 1994. DOI: 10.1016/0959-4752(94)90003-5. Acesso em: 20 abr. 2025.

SWELLER, J. Element interactivity and intrinsic, extraneous, and germane cognitive load. **Educational Psychology Review**, v. 22, n. 2, p. 123-138, 2010. DOI: 10.1007/s10648-010-9128-5. Acesso em: 20 abr. 2025.

SWELLER, J.; CHANDLER, P. Why some material is difficult to learn. **Cognition and Instruction**, v. 12, n. 3, p. 185-233, 1994. DOI: 10.1207/s1532690xci1203_1. Acesso em: 20 abr. 2025.

SWELLER, J.; VAN MERRIENBOER, J. J. G.; PAAS, F. Cognitive Architecture and Instructional Design. **Educational Psychology Review**, v. 10, n. 3, p. 251-296, 1998. DOI: 10.1023/A:1022193728205. Acesso em: 20 abr. 2025.

TANTITHAMTHAVORN, C. *et al*. MLOps, LLMOps, FMOps, and Beyond. **IEEE Software**, v. 42, n. 1, p. 26-32, jan./fev. 2025. DOI: 10.1109/MS.2024.3477014.

TATE, T. *et al*. **ChatGPT and the Future of University Assessment**. [S. l.]: EdArXiv, 2023. DOI: 10.35542/osf.io/562k3. Disponível em: https://edarxiv.org/562k3/. Acesso em: 25 abr. 2025.

THOMAS, J. W. **A review of research on project-based learning**. San Rafael, CA: Autodesk Foundation, 2000. Disponível em: http://www.bie.org/research/study/review_of_project_based_learning. Acesso em: 25 abr. 2025.

TRUONG, Q.-T. *et al*. **Towards Automated Testing of LLM-based Applications**: A Survey. [S. l.]: arXiv, 2023. Disponível em: https://arxiv.org/abs/2311.18119. Acesso em: 20 abr. 2025.

VAN MERRIENBOER, J. J. G.; SWELLER, J. Cognitive load theory and complex learning: Recent developments and future directions. **Educational Psychology Review**, v. 17, n. 2, p. 147-177, 2005. DOI: 10.1007/s10648-005-3951-0. Acesso em: 20 abr. 2025.

WANG, H. *et al*. **Leveraging Reasoning Model Answers to Enhance Non-Reasoning Model Capability**. [S. l.]: arXiv, 2025a. Disponível em: https://arxiv.org/abs/2504.09639. Acesso em: 20 abr. 2025.

WANG, Hongru *et al*. **OTC**: Optimal Tool Calls via Reinforcement Learning. [S. l.]: arXiv, 21 abr. 2025b. Disponível em: https://arxiv.org/html/2504.14870v1. Acesso em: 5 maio 2025.

WOOD, D.; BRUNER, J. S.; ROSS, G. The role of tutoring in problem solving. **Journal of Child Psychology and Psychiatry**, v. 17, n. 2, p. 89-100, 1976. DOI: 10.1111/j.1469-7610.1976.tb00381.x. Acesso em: 20 abr. 2025.

WU, T. *et al*. **AutoGen**: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. [S. l.]: arXiv, 2023. Disponível em: https://arxiv.org/abs/2308.08155. Acesso em: 20 abr. 2025.

XIA, S. *et al*. **Unlocking Deep Thinking in Language Models**: Cognition Engineering through Inference Time Scaling and Reinforcement Learning. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2504.13828. Acesso em: 24 abr. 2025.

XINOGALOS, S. Cognitive Load in Computer Science Education: A Systematic Literature Review. **IEEE Transactions on Learning Technologies**, v. 14, n. 5, p. 648-664, Sept.-Oct. 2021. DOI: 10.1109/TLT.2021.3122001.

XU, W. **Human-Centered AI**. Boca Raton: CRC Press, 2019.

YAO, Z. *et al*. **ZeroQuant**: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers. [S. l.]: arXiv, 2022. Disponível em: https://arxiv.org/abs/2206.01861. Acesso em: 20 abr. 2025.

YETISTIREN, B. *et al*. **A Systematic Literature Review on the Use of Large Language Models in Software Engineering**. [S. l.]: arXiv, 2024. Disponível em: https://arxiv.org/abs/2402.18551. Acesso em: 25 abr. 2025.

ZHANG, J.; NORMAN, D. A. Representations in distributed cognitive tasks. **Cognitive Science**, v. 18, n. 1, p. 87-122, 1994. DOI: 10.1207/s15516709cog1801_3. Acesso em: 20 abr. 2025.

ZHANG, T. *et al*. **LLM-Orchestrator**: Orchestrating Collaborative Large Language Models for Automated Software Engineering. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2402.16831. Acesso em: 20 abr. 2025.

ZHAO, P.; YAUN, X. **GANQ**: GPU-Adaptive Non-Uniform Quantization for Large Language Models. [S. l.]: arXiv, 2025. Disponível em: https://arxiv.org/abs/2501.12956. Acesso em: 20 abr. 2025.

ZIMMERMAN, B. J. Becoming a self-regulated learner: An overview. **Theory Into Practice**, v. 41, n. 2, p. 64-70, 2002. DOI: 10.1207/s15430421tip4102_2. Acesso em: 20 abr. 2025.