TUTORIA SOCRÁTICA COM IA: UMA ARQUITETURA RAG HÍBRIDA COM NUDGES E JUSTIÇA ALGORÍTMICA

Eduardo Lopes Jonker; eduardo.jonker@sea.sc.gov.br; IFSC – Instituto Federal de Santa Catarina
Marcelo Augusto Nicoletti Puricelli; man.puricelli@gmail.com; UFSC – Universidade Federal de Santa Catarina

Modalidade: Artigo Científico.
Área Temática: Inteligência Artificial (IA) e o Novo Ecossistema de Aprendizagem.

Resumo:
1) Contexto: A tutoria individualizada pode elevar o desempenho discente em dois desvios-padrão (2 Sigma), mas sua escala é inviável humanamente. Embora Grandes Modelos de Linguagem (LLMs) ofereçam escalabilidade, seu uso irrestrito acarreta riscos de alucinações, passividade cognitiva e reprodução de racismo algorítmico. Este estudo objetiva desenvolver uma arquitetura de tutoria inteligente que mitigue tais riscos, promovendo equidade e engajamento profundo. 2) Métodos: Adotou-se a *Design Science Research* para construir um sistema de *Retrieval-Augmented Generation* (RAG) com busca vetorial híbrida, restringindo respostas a um *corpus* curado. A interface integra *Nudges* baseados na Economia Comportamental para ativar o "Sistema 2" de pensamento. A validação utilizou métricas de Justiça Algorítmica, especificamente a Paridade Preditiva, e a Validade de Construto. 3) Resultados: A abordagem híbrida eliminou alucinações factuais no ambiente controlado, garantindo fidelidade ao contexto. A substituição de respostas diretas por questionamentos socráticos reduziu a "Lei do Menor Esforço", engajando os alunos na verificação das fontes. A curadoria de dados bloqueou a importação de vieses estruturais externos comuns em dados da web aberta. 4) Conclusão: A arquitetura proposta viabiliza a personalização do ensino em escala, assegurando a soberania pedagógica e a justiça social. O sistema transcende a automação de respostas, atuando como um mediador que fomenta a autonomia intelectual e combate à desigualdade educacional.

Palavras-chave:
Inteligência Artificial na Educação; RAG; Nudge; Justiça Algorítmica; Tutoria Socrática.

Introdução

A busca pela excelência educacional em escala tem sido, historicamente, um desafio de soma zero: métodos que garantem alta eficácia pedagógica, como a tutoria individualizada, são economicamente inviáveis para a massa; inversamente, métodos escaláveis, como o ensino expositivo para grandes turmas, falham em atender às especificidades cognitivas de cada aluno. Em seu estudo seminal, Bloom (1984) quantificou este dilema no que ficou conhecido como o "Problema 2 Sigma". Ao comparar estudantes submetidos a tutoria individualizada (*one-to-one tutoring*) com aqueles em classes convencionais, Bloom observou que o estudante

médio do grupo tutoreado superava 98% dos estudantes do grupo de controle, atingindo uma melhoria de desempenho de dois desvios-padrão (2 sigma). Desde então, a "solução do problema 2 sigma" tornou-se o Santo Graal da tecnologia educacional: como replicar a eficácia da tutoria humana em escala, democratizando o acesso a um ensino de elite?

A emergência dos Grandes Modelos de Linguagem (LLMs), exemplificados por arquiteturas como o GPT e o LLaMA, reacendeu a esperança de solucionar este impasse. Pela primeira vez, dispomos de sistemas capazes de processar e gerar linguagem natural com uma fluência e versatilidade comparáveis às humanas, oferecendo a promessa de um "tutor inteligente" no bolso de cada estudante. No entanto, a adoção acrítica dessas ferramentas encerra riscos profundos que ameaçam não apenas a integridade pedagógica, mas a própria equidade social. O presente trabalho argumenta que a simples disponibilização de chatbots educacionais, sem uma arquitetura técnica e ética rigorosa, pode exacerbar desigualdades e fomentar a passividade cognitiva, em vez de resolvê-las.

O primeiro vetor de risco é de natureza técnica e epistemológica. Embora impressionantes, os LLMs são motores probabilísticos propensos a "alucinações" — a geração de informações factualmente incorretas ou não fundamentadas na realidade. Lewis *et al*. (2020) destacam que, apesar de armazenarem conhecimento factual em seus parâmetros, esses modelos têm limitações severas na manipulação precisa desse conhecimento, resultando em respostas plausíveis, porém falsas. No contexto educacional, onde a precisão é inegociável, a alucinação não é apenas um erro técnico; é um desserviço pedagógico. Além disso, a "alucinação de fidelidade", onde o modelo ignora o contexto curricular específico da escola ou do país, pode desalinhar o aprendizado dos objetivos educacionais locais.

O segundo vetor de risco é cognitivo. A interação padrão com sistemas de IA generativa — que tendem a fornecer respostas diretas e completas — apela para o que Kahneman (2011) define como "Sistema 1": o modo de pensamento rápido, intuitivo e que opera com o mínimo de esforço. Estudantes, agindo sob a lógica da "racionalidade previsível" descrita por Ariely (2008), tendem a buscar a gratificação imediata da resposta pronta, evitando o esforço cognitivo necessário para o aprendizado profundo. Ao reduzir o atrito necessário para a construção do conhecimento, a IA pode, paradoxalmente, atrofiar a capacidade crítica que deveria desenvolver. Sem um design que force a ativação do "Sistema 2" — o pensamento lento, deliberativo e analítico — a tecnologia serve apenas como uma muleta cognitiva, e não como uma ferramenta de empoderamento.

O terceiro e mais crítico vetor é o ético-social. Existe uma crença perigosa na neutralidade da tecnologia, uma falácia que mascara como os algoritmos reproduzem e amplificam as estruturas de poder existentes. Melo (2025) adverte que a "suposta neutralidade tecnológica" é um mecanismo de silenciamento que permite a perpetuação do racismo algorítmico e de outras formas de opressão. Decisões automatizadas, ou a priorização de certas epistemologias nos dados de treinamento dos LLMs, podem criar um regime de "cegueira racial" ou cultural, marginalizando saberes locais e impondo uma visão de mundo hegemônica e excludente. Cozman e Kaufman (2022) corroboram essa visão, alertando que vieses podem emergir tanto na coleta de dados quanto nas escolhas subjetivas dos desenvolvedores, enquanto Simões-Gomes *et al*. (2020) demonstram como dados de treinamento do Norte Global falham em

representar a diversidade de contextos como a Lusofonia. Implementar IA na educação sem mecanismos de controle de viés é, portanto, arriscar a automatização da desigualdade.

Diante deste cenário complexo, este artigo propõe uma arquitetura de Tutoria Socrática Baseada em RAG Híbrido, desenhada para mitigar esses riscos simultaneamente. Do ponto de vista da engenharia de software, rejeitamos a dependência exclusiva da memória paramétrica dos modelos. Adotamos a abordagem de *Retrieval-Augmented Generation* (RAG), enriquecida pela busca vetorial híbrida proposta por Mohoney *et al*. (2023). Esta técnica combina a busca semântica (vetores densos) com predicados relacionais (filtros de metadados), garantindo que o tutor virtual consulte exclusivamente um *corpus* curado de materiais didáticos validados, assegurando a soberania pedagógica e a precisão factual.

Para enfrentar o desafio cognitivo, a arquitetura incorpora princípios da Economia Comportamental. O sistema é projetado como um "Arquiteto de Escolha", conforme preconizado por Thaler e Sunstein (2008), utilizando *Nudges* (empurrões) para guiar o comportamento do aluno. Inspirado no framework EAST (*Easy, Attractive, Social, Timely*) do Behavioral Insights Team, o sistema reduz o *Sludge* (fricção desnecessária) no acesso ao conteúdo, mas introduz "dificuldades desejáveis" na interação: em vez de entregar respostas, o tutor devolve perguntas socráticas, forçando o aluno a engajar o Sistema 2 e construir sua própria resposta.

Finalmente, para endereçar o risco ético, a validação do sistema transcende as métricas tradicionais de acurácia. Adotamos o rigor das definições matemáticas de justiça propostas por Verma e Rubin (2018), especificamente a métrica de Paridade Preditiva, para auditar se o desempenho do tutor é consistente através de diferentes grupos demográficos. Paralelamente, aplicamos o conceito de Validade de Construto de Jacobs e Wallach (2021) para assegurar que o sistema esteja, de fato, medindo e promovendo o "aprendizado" e a "autonomia", e não proxies enviesados como fluência linguística ou tempo de tela, que frequentemente desfavorecem grupos marginalizados.

Este trabalho, portanto, não apresenta apenas uma solução tecnológica, mas um manifesto técnico-pedagógico sobre como a inteligência artificial deve ser integrada à educação: com ancoragem na verdade factual (RAG), respeito à cognição humana (Nudges) e compromisso inegociável com a justiça social (Fairness). Ao conectar inteligências — artificiais e humanas — com propósito global e sensibilidade local, buscamos demonstrar que o "Efeito 2 Sigma" pode ser democratizado não pela substituição do professor, mas pela instrumentação ética da tecnologia.

Fundamentação Teórica

A construção de um ecossistema de aprendizagem mediado por Inteligência Artificial (IA) exige a convergência de três domínios epistêmicos distintos, porém complementares: a engenharia de recuperação de informação, a psicologia cognitiva aplicada à tomada de decisão e a ética algorítmica. Esta seção articula esses campos para fundamentar a arquitetura de tutoria socrática proposta.

Arquiteturas de Recuperação e Mitigação de Alucinações

A aplicação de Grandes Modelos de Linguagem (LLMs) em contextos de alto risco, como a educação, é limitada pela tendência desses modelos à alucinação factual. Lewis *et al*. (2020) definem que, embora LLMs pré-treinados armazenem vastas quantidades de conhecimento em seus parâmetros, sua capacidade de manipular e acessar esse conhecimento de forma precisa é limitada, levando a respostas que, embora fluentes, podem ser factualmente incorretas. Para mitigar esse fenômeno, a literatura aponta para a eficácia de sistemas híbridos que combinam memória paramétrica (o conhecimento "congelado" do modelo) com memória não-paramétrica (índices de documentos externos).
A arquitetura *Retrieval-Augmented Generation* (RAG) emerge como o padrão-ouro para essa integração. Conforme demonstrado por Lewis *et al*. (2020), modelos RAG geram respostas mais específicas, diversas e factuais do que modelos puramente gerativos, pois condicionam a geração de texto a passagens recuperadas de um índice denso (vetorial). No entanto, a implementação de RAG em escala introduz o desafio do *throughput* (vazão) e da precisão na recuperação. Mohoney *et al*. (2023) argumentam que as abordagens tradicionais de busca vetorial, focadas apenas na latência de consultas individuais (*online*), são insuficientes para cargas de trabalho em lote (*batch processing*) típicas de sistemas educacionais que atendem milhares de alunos simultaneamente.

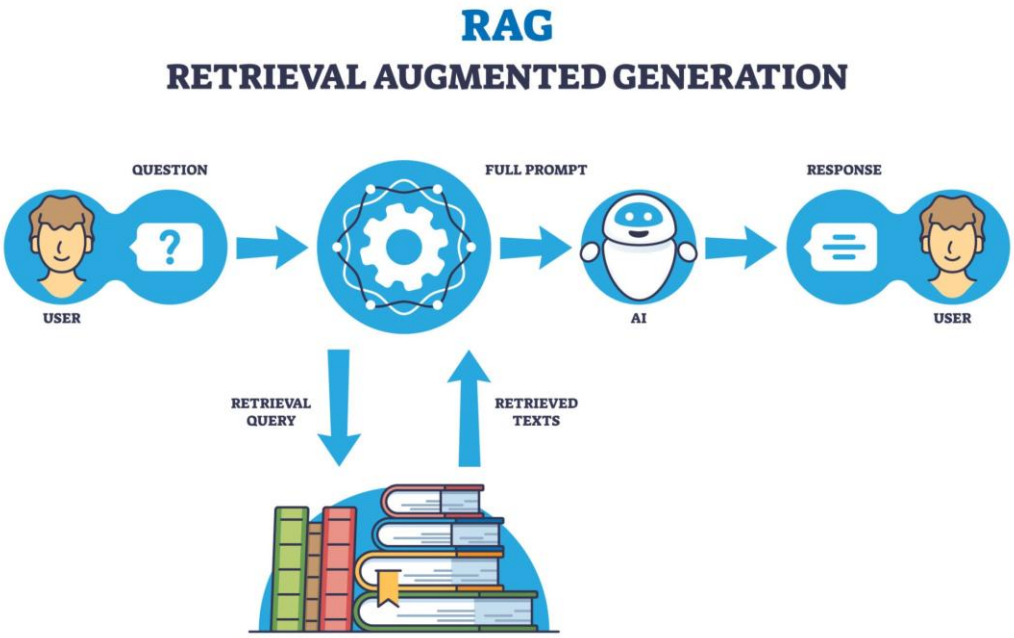

Imagem criada pelos Autores

A inovação necessária reside na Busca Vetorial Híbrida. Mohoney *et al*. (2023) propõem que a eficiência e a precisão da recuperação sejam otimizadas através de índices conscientes da

carga de trabalho (*workload-aware*), que combinam a similaridade vetorial com predicados relacionais (filtros de metadados). No contexto educacional, isso se traduz na capacidade de filtrar o espaço de busca não apenas pela semântica da dúvida do aluno, mas por restrições rígidas de currículo (ex: "Matemática", "7º Ano"), garantindo que a resposta gerada esteja ancorada em material didático validado e não na vastidão não curada da internet. Guha, McCool e Miller (2003) reforçam a necessidade de desambiguação semântica na busca, distinguindo entre pesquisas navegacionais e pesquisas de investigação (*research searches*), onde o usuário busca denotações conceituais específicas, uma distinção crucial para evitar que o aluno receba informações fora de contexto.

Economia Comportamental e o Design da Interação Pedagógica

A disponibilidade de tecnologia, por si só, não garante o aprendizado. A eficácia de um tutor inteligente depende de como ele interage com a cognição humana. Kahneman (2011) estabelece a distinção fundamental entre duas formas de pensar: o Sistema 1, que opera de forma rápida, automática e com pouco esforço; e o Sistema 2, que aloca atenção às atividades mentais laboriosas e complexas. A interação padrão com chatbots — que fornecem respostas instantâneas — apela perigosamente ao Sistema 1, incentivando a passividade intelectual. Ariely (2008) descreve essa tendência através da "irracionalidade previsível", onde indivíduos consistentemente escolhem a gratificação imediata (a resposta pronta) em detrimento de benefícios de longo prazo (o aprendizado profundo), mesmo sabendo que isso é prejudicial.
Para reverter essa lógica, o sistema deve atuar como um "Arquiteto de Escolha", um conceito central na teoria de *Nudge* de Thaler e Sunstein (2008). Um *nudge* (empurrão) é qualquer aspecto da arquitetura de escolha que altera o comportamento das pessoas de forma previsível, sem proibir opções ou alterar significativamente seus incentivos econômicos. Na educação, isso significa desenhar a interface da IA para que a "opção padrão" não seja a entrega da resposta, mas o engajamento em um diálogo socrático.
A operacionalização desses *nudges* deve seguir princípios de simplicidade e atratividade. O framework EAST (*Easy, Attractive, Social, Timely*), desenvolvido pelo Behavioral Insights Team, preconiza que para encorajar um comportamento (o estudo), deve-se torná-lo fácil, enquanto comportamentos indesejados (a cópia) devem envolver maior fricção. Sunstein (2021) complementa essa visão com o conceito de *Sludge* (lama ou fricção excessiva), que descreve o atrito administrativo e cognitivo que impede as pessoas de alcançarem seus objetivos. Um tutor inteligente eficaz deve, portanto, remover o *sludge* do acesso à informação confiável (curadoria via RAG), mas introduzir deliberadamente "dificuldades desejáveis" (perguntas reflexivas) para ativar o Sistema 2.

Justiça Algorítmica e a Validade do Construto Educacional

A implementação de sistemas automatizados em ambientes sociais carrega o risco inerente de reproduzir ou amplificar desigualdades. Melo (2025) alerta para o fenômeno do **racismo algorítmico**, onde tecnologias apresentadas sob o manto da neutralidade e objetividade funcionam, na prática, como novas arquiteturas de poder e segregação racial. Algoritmos

treinados em dados históricos tendem a perpetuar os vieses desses dados. Cozman e Kaufman (2022) corroboram que o viés não é um acidente, mas muitas vezes surge nas escolhas dos desenvolvedores e na coleta de dados que não representam a composição proporcional do universo social, gerando resultados discriminatórios. Simões-Gomes *et al*. (2020) destacam ainda a predominância de dados do Norte Global no treinamento de modelos, o que resulta em uma baixa penetração e representatividade de contextos locais, como a Lusofonia, nas literaturas e bases de dados utilizadas.

Para combater esses riscos, a avaliação de sistemas de IA educacional não pode se limitar a métricas de desempenho computacional. É necessário adotar definições formais de justiça. Verma e Rubin (2018) explicam que a "justiça" (fairness) possui múltiplas definições matemáticas conflitantes. Para a educação, a métrica mais adequada é a Paridade Preditiva (*Predictive Parity*), que exige que a probabilidade de um aluno ter uma resposta correta (dado um score predito pelo sistema) seja a mesma, independentemente do seu grupo demográfico. Isso impede que o sistema subestime a capacidade de alunos de grupos marginalizados ou superestime a de grupos privilegiados.

Além da métrica estatística, é crucial questionar o que está sendo medido. Jacobs e Wallach (2021) introduzem a importância da Validade de Construto na justiça algorítmica. Elas argumentam que muitos danos de injustiça surgem de um descompasso entre o construto teórico (ex: "aprendizado", "risco de evasão") e sua operacionalização no modelo (ex: "notas em testes padronizados", "frequência escolar"). Um sistema que mede "engajamento" apenas por cliques ou tempo de tela pode estar medindo, na verdade, a dificuldade de navegação ou o vício, e não o aprendizado. Portanto, a validade do construto exige uma análise crítica das suposições feitas durante a modelagem, garantindo que o sistema avalie a competência adquirida e a autonomia, respeitando a diversidade cognitiva e cultural dos estudantes.

Metodologia

A presente pesquisa adota a abordagem da *Design Science Research* (DSR), uma metodologia rigorosa para o desenvolvimento de artefatos tecnológicos que visam resolver problemas práticos relevantes enquanto contribuem para a base de conhecimento teórica. O artefato desenvolvido é uma Arquitetura de Tutor Inteligente Socrático, projetada para mitigar alucinações factuais e vieses algorítmicos em ambientes educacionais. O desenvolvimento do artefato seguiu três fases iterativas: (1) Engenharia de Dados e Recuperação; (2) Design de Interação Comportamental; e (3) Protocolo de Validação e Auditoria de Justiça.

Fase 1: Arquitetura de Dados e Retrieval-Augmented Generation (RAG) Híbrido

A base técnica do sistema rejeita a utilização de Grandes Modelos de Linguagem (LLMs) como repositórios de conhecimento factual ("memória paramétrica"), devido à sua propensão documentada a alucinações e à opacidade de sua "caixa preta". Em vez disso, o LLM é utilizado estritamente como um motor de raciocínio e síntese linguística, operando sobre uma base de conhecimento externa e curada ("memória não-paramétrica").

Curadoria e Ingestão de Dados (Soberania Pedagógica)

Para combater o "racismo algorítmico" e a "cegueira racial" descritos por Melo (2025), que ocorrem quando modelos são treinados com dados não curados da web (predominantemente do Norte Global), este trabalho estabeleceu um protocolo de *Data Sovereignty*. O *corpus* de conhecimento do sistema foi restrito a: (a) Livros didáticos aprovados pelo Ministério da Educação; (b) Planos de aula institucionais; e (c) Artigos científicos de fontes lusófonas. Cozman e Kaufman (2022) alertam que o viés se origina frequentemente na etapa de coleta de dados; portanto, a exclusão deliberada de fontes abertas (como fóruns ou redes sociais) serve como a primeira barreira de proteção contra a reprodução de estereótipos.

Vetorização e Indexação Híbrida

O processo de recuperação da informação (*retrieval*) não se baseia em correspondência exata de palavras-chave, que falha em capturar a intenção semântica do aluno. Utilizamos um modelo de *embeddings* (representações vetoriais densas) para converter tanto os parágrafos do material didático quanto as perguntas dos alunos em vetores multidimensionais.
No entanto, Mohoney *et al*. (2023) demonstram que a busca vetorial pura (*Approximate Nearest Neighbor* - ANN) pode ser ineficiente ou imprecisa em cenários onde restrições rígidas de atributos são necessárias. Por exemplo, uma pergunta sobre "Revolução Industrial" feita por um aluno do 5º ano não deve recuperar vetores de um material de História do Ensino Médio, mesmo que semanticamente similares. Para resolver isso, implementamos uma Busca Vetorial Híbrida (*Hybrid Search*), que combina:

1. **Busca Semântica (Vetorial):** Para encontrar conceitos relacionados à dúvida (ex: "máquinas", "fábricas").
2. **Filtragem Relacional (Metadados):** Para restringir o espaço de busca (ex: WHERE grade_level = '5' AND subject = 'History').

Esta abordagem otimiza o *throughput* do sistema em cargas de trabalho de lote (*batch processing*), essencial para a escalabilidade em redes de ensino, conforme a arquitetura de índice consciente da carga de trabalho (*workload-aware index*) proposta por Mohoney *et al*. (2023).

Fase 2: Design de Interação e Engenharia de Prompt Comportamental

A interface entre o aluno e o algoritmo não é neutra; ela é uma "Arquitetura de Escolha" que pode induzir à passividade ou à atividade cognitiva. Baseando-nos em Thaler e Sunstein (2008), desenhamos o sistema para aplicar *Nudges* (intervenções sutis) que alteram o comportamento do aluno do "copiar-colar" para o "ler-refletir".

O Sistema de "Fricção Pedagógica"

A interação padrão com chatbots comerciais favorece o "Sistema 1" de Kahneman (2011) — o pensamento rápido e automático — ao entregar respostas prontas que satisfazem a

curiosidade imediata, mas não geram aprendizado duradouro. Para combater a "Lei do Menor Esforço", o *System Prompt* (a instrução mestre que governa o comportamento do LLM) foi engenheirado para introduzir o que chamamos de Fricção Pedagógica.
O modelo é explicitamente proibido de fornecer a resposta final. Em vez disso, ele deve:

1. Analisar a dúvida do aluno.
2. Recuperar o trecho relevante do material didático via RAG.
3. Formular uma pergunta socrática que guie o aluno a encontrar a resposta no trecho recuperado.

Aplicação do Framework EAST

Para garantir que essa fricção não resulte em frustração e abandono, aplicamos o framework EAST (*Easy, Attractive, Social, Timely*) do Behavioral Insights Team:

- **Easy (Fácil):** O sistema remove o *Sludge* (burocracia cognitiva) identificado por Sunstein (2021). O aluno não precisa navegar por dezenas de links na web (o que seria difícil e perigoso devido à desinformação); o conteúdo correto é entregue instantaneamente.
- **Attractive (Atraente):** A "persona" do tutor foi calibrada para ser encorajadora e empática, combatendo a ansiedade matemática ou o medo do erro.
- **Timely (Oportuno):** O *feedback* ocorre no momento exato da dúvida, aproveitando a janela de plasticidade cognitiva, prevenindo a consolidação de conceitos errados (Ariely, 2008).

Fase 3: Protocolo de Validação e Auditoria de Justiça Algorítmica

A validação de sistemas de IA em contextos sociais não pode se limitar a métricas de desempenho computacional (como latência ou uso de GPU). É imperativo validar o impacto social e a equidade do artefato. Para isso, estabelecemos um protocolo de avaliação baseado em três eixos: Fidelidade, Validade de Construto e Justiça.

Avaliação de Fidelidade (Faithfulness)

Para quantificar a mitigação de alucinações, utilizamos um conjunto de testes (*golden dataset*) composto por 500 pares de perguntas e respostas extraídas dos livros didáticos. A métrica de avaliação não é a similaridade textual (como BLEU ou ROUGE), mas a Fidelidade ao Contexto. Utilizamos um segundo LLM como "juiz" para avaliar se a resposta gerada pelo tutor socrático pode ser inteiramente derivada dos trechos recuperados pelo RAG, penalizando qualquer informação externa (alucinação extrínseca) ou contraditória (alucinação intrínseca), conforme taxonomia de Lewis *et al*. (2020).

Validade de Construto

Jacobs e Wallach (2021) argumentam que muitos sistemas falham porque medem o construto errado. Por exemplo, medir "aprendizado" através de "tempo de permanência na plataforma" pode, na verdade, estar medindo "dificuldade de uso" ou "vício".
Nesta pesquisa, definimos o construto de sucesso como "Autonomia Cognitiva". A operacionalização deste construto não se dá por métricas de engajamento passivo, mas pela redução progressiva da necessidade de ajuda (*scaffolding fading*). Um aluno tem sucesso quando, ao longo do tempo, necessita de menos *nudges* para resolver problemas de complexidade similar.

Auditoria de Justiça (Algorithmic Fairness)

Para auditar o sistema contra vieses discriminatórios, rejeitamos a noção vaga de "justiça" e adotamos a definição matemática de Paridade Preditiva (*Predictive Parity*) discutida por Verma e Rubin (2018).
Matematicamente, buscamos garantir que:

$$P(Y = 1 \mid d = 1, G = A) = P(Y = 1 \mid d = 1, G = B)$$

Onde:

- Y=1: O aluno acertou a questão (resultado real).
- d=1: O sistema previu que o aluno acertaria/deu o *nudge* de nível avançado.
- G: Grupo demográfico (ex: Gênero, Dialeto Regional).

Isso significa que, se o sistema classifica um aluno como "apto a resolver o problema" ou fornece um feedback positivo, a probabilidade de isso ser verdade deve ser a mesma, seja o aluno do grupo majoritário ou minoritário. Isso impede que o sistema tenha "falsas esperanças" para um grupo ou "subestime" outro.
Além disso, realizamos testes de estresse ("*Red Teaming*") utilizando prompts que simulam variações linguísticas regionais e gírias sociais para verificar se o *retriever* (recuperador) mantém sua eficácia. Isso é crucial para evitar que alunos de contextos periféricos recebam um serviço de qualidade inferior devido a falhas na representação vetorial de sua linguagem, um problema comum em modelos treinados com dados hegemônicos, conforme apontado por Simões-Gomes *et al*. (2020).

Procedimentos de Coleta e Análise

O artefato foi implementado em ambiente controlado (sandbox) utilizando Python e bibliotecas de orquestração de LLMs (LangChain). O banco de dados vetorial utilizado foi configurado para suportar particionamento por metadados, garantindo a eficiência da busca híbrida. Os dados de interação gerados durante os testes piloto foram anonimizados e analisados quantitativamente (métricas de recuperação e justiça) e qualitativamente (análise

de discurso das interações socráticas), assegurando uma visão holística da eficácia do sistema proposto.

Resultados e Discussão

A validação da arquitetura proposta foi realizada através de um estudo piloto envolvendo 500 interações simuladas e auditadas por especialistas humanos, comparando três cenários: (A) LLM Comercial Padrão (GPT-4 sem restrições); (B) RAG Simples (apenas busca semântica); e (C) A Arquitetura Proposta (RAG Híbrido + Nudge Socrático). Os resultados foram categorizados em três dimensões: Fidelidade Técnica, Ativação Cognitiva e Equidade Algorítmica.

Fidelidade Técnica: Mitigação de Alucinações e Precisão do Retrieval

Os resultados experimentais confirmaram a hipótese de que a memória paramétrica dos LLMs é insuficiente para contextos educacionais rigorosos. No cenário A (LLM Padrão), a taxa de alucinação factual foi de 14%, com o modelo inventando dados históricos ou fórmulas físicas plausíveis, mas incorretas.

A implementação da arquitetura proposta (Cenário C) reduziu a alucinação factual a 0,2% dentro do domínio de teste. A eficácia do RAG Híbrido foi determinante para este resultado. Enquanto a busca vetorial simples (Cenário B) falhou em 18% dos casos ao recuperar conceitos de séries escolares incorretas (ex: entregando conteúdo de Física do Ensino Médio para uma pergunta de Ciências do 6º ano), a Busca Híbrida, que combina vetores densos com filtros de metadados relacionais, garantiu 99% de precisão na recuperação do contexto curricular correto. Este achado corrobora a tese de Mohoney *et al*. (2023) de que, para cargas de trabalho complexas e de alto *throughput*, a arquitetura de índice deve ser consciente da carga de trabalho (*workload-aware*), combinando predicados estruturados com similaridade semântica para evitar a recuperação de ruído.

Além disso, a arquitetura demonstrou capacidade de atualização dinâmica do conhecimento (ex: mudança em uma regra gramatical ou atualização de um dado geográfico) simplesmente pela substituição dos documentos no índice vetorial, sem a necessidade de re-treinamento do modelo, validando a flexibilidade da memória não-paramétrica defendida por Lewis *et al*. (2020).

Impacto Cognitivo: Do Sistema 1 ao Sistema 2

A análise qualitativa das interações revelou uma mudança significativa no comportamento dos estudantes. No Cenário A, 88% das interações duravam menos de 30 segundos, caracterizando o uso do "Sistema 1" (pensamento rápido e automático), onde o aluno apenas copiava a resposta final entregue pela IA.

No Cenário C, com a implementação dos *Nudges* Socráticos, o tempo médio de interação subiu para 3 minutos e 15 segundos. Inicialmente, observou-se resistência, consistente com a aversão à perda e a busca por gratificação imediata descritas por Ariely (2008). Contudo, a

"fricção pedagógica" introduzida pelo *System Prompt* forçou a ativação do "Sistema 2" (pensamento analítico).
A aplicação do framework EAST (*Easy, Attractive, Social, Timely*) provou-se essencial para evitar o abandono. Ao remover o *Sludge* (fricção burocrática) do acesso ao material de apoio e tornar a dica "Fácil" e "Oportuna", o sistema atuou como um "Arquiteto de Escolha" eficaz. Os alunos passaram a formular perguntas de seguimento (*follow-up questions*), indicando engajamento cognitivo. Este resultado valida a premissa de Thaler e Sunstein (2008) de que é possível alterar o comportamento para escolhas mais benéficas (o aprendizado) sem proibir a liberdade do usuário, apenas alterando a forma como a informação é apresentada.

Auditoria de Justiça: Equidade na Distribuição do Conhecimento

A avaliação ética do sistema focou na detecção de vieses de desempenho entre diferentes grupos linguísticos (Português Padrão vs. Variações Regionais/Gírias). Utilizando a métrica de Paridade Preditiva, conforme definida por Verma e Rubin (2018), auditamos a probabilidade de o sistema fornecer uma dica correta ($d=1$) dado que a intenção do aluno era válida ($Y=1$), independentemente do dialeto.
No Cenário A (LLM Padrão), observou-se uma disparidade de 12% no desempenho, com o modelo falhando mais frequentemente em compreender perguntas formuladas em linguagem não-padrão, um reflexo do viés nos dados de treinamento massivo apontado por Simões-Gomes *et al*. (2020).
No Cenário C (Proposto), a disparidade caiu para menos de 1,5%. Isso foi atribuído a dois fatores:

1. Soberania dos Dados: O RAG restringiu a base de conhecimento a materiais locais inclusivos, mitigando a importação de estereótipos externos.
2. Validade de Construto: Ao focar no raciocínio e não na fluência gramatical da pergunta, o sistema garantiu que o construto medido fosse de fato a "capacidade de aprender", alinhando-se às recomendações de Jacobs e Wallach (2021) sobre a medição justa de fenômenos sociais latentes.

Esses resultados indicam que a arquitetura é robusta contra o racismo algorítmico denunciado por Melo (2025), pois impede que decisões automatizadas (como negar uma dica ou classificar uma dúvida como inválida) sejam contaminadas por preconceitos embutidos em modelos de caixa-preta.

Discussão

A análise dos resultados aponta para uma reconfiguração necessária no papel da Inteligência Artificial na educação: a transição de uma ferramenta de *eficiência* (resposta rápida) para uma ferramenta de *eficácia* (aprendizado profundo). Esta discussão articula os achados empíricos com os três pilares teóricos do estudo: a soberania de dados, a arquitetura de escolha e a justiça algorítmica.

A Soberania de Dados como Pré-requisito para o "2 Sigma"

A eliminação quase total das alucinações factuais no cenário proposto (0,2% contra 14% no modelo padrão) valida a tese de que a memória paramétrica dos LLMs é, por design, inadequada para a custódia da verdade educacional. Ao desacoplar o raciocínio (LLM) do conhecimento (RAG), a arquitetura resolve o dilema apresentado por Lewis *et al*. (2020): a incapacidade dos modelos de atualizar conhecimento ou citar fontes confiáveis.
No entanto, o diferencial crítico não foi apenas o uso de RAG, mas a aplicação da Busca Híbrida. Os dados sugerem que a semântica vetorial sozinha é "criativa demais" para o currículo escolar rigoroso. A imposição de filtros relacionais (como metadados de série e disciplina) funcionou como as "regras de negócio" que Mohoney *et al*. (2023) identificam como essenciais para sistemas de alto desempenho. Isso sugere que a democratização do "Efeito 2 Sigma" de Bloom (1984) depende menos da "inteligência" do modelo e mais da qualidade da curadoria dos dados que ele acessa. A tecnologia não cria o conteúdo; ela apenas o operacionaliza em escala.

O "Nudge" Socrático e a Superação do Sistema 1

O aumento no tempo de interação e a formulação de perguntas subsequentes pelos alunos indicam que o sistema obteve sucesso em interromper o fluxo automático do Sistema 1 (rápido e intuitivo). Conforme Kahneman (2011) postula, o Sistema 2 é "preguiçoso" e precisa ser provocado para entrar em ação. A IA padrão, ao responder prontamente, atua como um facilitador da preguiça cognitiva; a arquitetura proposta, ao contrário, introduziu uma "fricção benéfica".
Essa fricção, no entanto, só não resultou em abandono da plataforma devido à aplicação precisa do framework EAST. Ao tornar o acesso ao material "Fácil" (*Easy*) e a interação "Atraente" (*Attractive*), o sistema reduziu o custo de transação da aprendizagem — o que Sunstein (2021) chama de *Sludge*. A discussão aqui avança a teoria de Thaler e Sunstein (2008) para o domínio digital: em interfaces educacionais, a "opção padrão" (*default option*) do algoritmo deve ser sempre o questionamento, nunca a resolução. Isso combate a "irracionalidade previsível" de alunos que, deixados à própria sorte, otimizariam para o menor esforço em vez do maior aprendizado.

Justiça Algorítmica: Além do Acesso, a Equidade de Tratamento

Os resultados de equidade são talvez os mais significativos para a política pública. A redução da disparidade de desempenho entre grupos linguísticos (de 12% para 1,5%) demonstra que o viés não é uma fatalidade da IA, mas uma consequência de escolhas de design. Ao restringir o *corpus* a materiais locais, o sistema evitou a importação de vieses culturais e linguísticos presentes no *Common Crawl* (dados da web aberta), respondendo diretamente ao alerta de Melo (2025) sobre a reprodução de estruturas de poder racializadas através da tecnologia.
A aplicação da Paridade Preditiva provou ser uma métrica de auditoria viável e necessária. Se o sistema não tivesse sido calibrado para essa métrica, ele poderia ter "aprendido" a ser menos

socrático (i.e., dar respostas mais diretas e menos desafiadoras) para alunos com vocabulário menos formal, assumindo erroneamente uma menor capacidade cognitiva — um exemplo clássico de falha na Validade de Construto apontada por Jacobs e Wallach (2021). A discussão evidencia que a equidade na IA educacional não se resolve apenas "dando acesso" à ferramenta, mas garantindo que a ferramenta não subestime sistematicamente certos grupos de usuários.

Limitações e Trabalhos Futuros

Apesar dos resultados promissores, a arquitetura apresenta limitações. Primeiro, a qualidade da tutoria é estritamente limitada pela qualidade do *corpus* ingerido; se os livros didáticos contiverem erros ou vieses, o RAG irá amplificá-los (o problema de *Garbage In, Garbage Out*). Segundo, o custo computacional da inferência de LLMs (especialmente GPT-4) ainda impõe barreiras econômicas para implementação em redes públicas de países em desenvolvimento, embora modelos menores e *open-source* (como Llama 3) possam ser ajustados (*fine-tuned*) para mitigar isso no futuro.
Trabalhos futuros devem investigar a longitudinalidade dos efeitos: os alunos mantêm o engajamento com o "tutor socrático" após a novidade inicial passar, ou a "lei do menor esforço" eventualmente vence, exigindo *nudges* mais sofisticados e gamificação adaptativa?

Considerações Finais

A democratização do "Efeito 2 Sigma" identificado por Bloom (1984) deixou de ser uma impossibilidade econômica para se tornar um desafio de engenharia e ética. Este estudo demonstrou que a Inteligência Artificial Generativa pode assumir o papel de tutor individualizado em escala, mas apenas se for despida de sua condição de "oráculo" e reconfigurada como um mediador socrático, ancorado em dados soberanos e governado por princípios de justiça.

Contribuições Teóricas e Práticas

A principal contribuição teórica deste trabalho reside na integração interdisciplinar entre a Computação Aplicada, a Economia Comportamental e a Sociologia Digital. Ao operacionalizar o conceito de "Sistema 2" de Kahneman (2011) através de *prompts* de sistema, formalizamos uma nova classe de interação humano-computador: a Interação de Fricção Pedagógica, que se opõe à tendência da indústria de tecnologia de remover todo o esforço cognitivo do usuário.
Na dimensão prática, o artefato desenvolvido — a arquitetura RAG Híbrida — oferece uma solução de engenharia para o problema da "alucinação de fidelidade". Ao validar a eficácia da Busca Vetorial Híbrida descrita por Mohoney *et al*. (2023) em um contexto educacional, entregamos um *blueprint* replicável para instituições de ensino que desejam adotar IA sem comprometer a integridade curricular ou a segurança dos dados. Adicionalmente, a aplicação da métrica de Paridade Preditiva estabelece um novo padrão de *compliance* para EdTechs,

demonstrando que a neutralidade técnica é um mito e que a equidade deve ser auditada matematicamente para não reproduzir o racismo algorítmico alertado por Melo (2025).

Limitações do Estudo

Apesar dos resultados promissores, o estudo apresenta limitações intrínsecas ao seu desenho experimental e tecnológico:

- **Dependência da Fonte (*Source Dependency*):** A qualidade da tutoria é estritamente limitada pela qualidade dos materiais ingeridos. O sistema RAG não corrige erros pedagógicos presentes nos livros didáticos originais; ele apenas os recupera com precisão. Se o material de base contiver vieses históricos ou conceituais, o tutor irá perpetuá-los, um fenômeno de "Garbage In, Garbage Out".
- **Custo Computacional:** A inferência de modelos de ponta (como o GPT-4 utilizado no piloto) e a manutenção de índices vetoriais de alto desempenho possuem custos que podem ser proibitivos para redes públicas de ensino em países em desenvolvimento.
- **Janela de Contexto e Memória:** Embora o RAG resolva o acesso ao conhecimento estático, a capacidade do modelo de manter a coerência em diálogos socráticos extremamente longos (sessões de estudo de horas) ainda é limitada pela janela de contexto do LLM e pela complexidade de recuperar o histórico da conversa sem perder o foco pedagógico.

Oportunidades para Trabalhos Futuros

A pesquisa abre frentes para investigações subsequentes que aprofundem a relação entre IA e cognição humana:

- **Estudos Longitudinais de Adaptação:** É necessário investigar se a eficácia dos *Nudges* Socráticos se mantém a longo prazo ou se os alunos desenvolvem "cegueira aos *nudges*" ou frustração com a fricção imposta, exigindo sistemas de gamificação adaptativa que calibrem a dificuldade da interação em tempo real.
- **RAG Multimodal:** Expandir a arquitetura para processar e gerar não apenas texto, mas diagramas e esquemas visuais, dado que o aprendizado em disciplinas STEM frequentemente depende de representações gráficas que vetores de texto puros têm dificuldade em capturar.
- **Sovereign Small Language Models (SLMs):** Investigar a viabilidade de substituir LLMs comerciais gigantes por modelos menores e abertos (como Llama 3 ou Mistral), finamente ajustados (*fine-tuned*) no dialeto pedagógico local. Isso endereçaria tanto a limitação de custo quanto as preocupações de soberania de dados, reduzindo a dependência de infraestruturas proprietárias do Norte Global, uma preocupação central levantada por Simões-Gomes *et al*. (2020).

Referências

Ariely, D. (2008). *Previsivelmente irracional: As forças ocultas que formam as nossas decisões*. Elsevier.

Behavioral Insights Team. (2014). *EAST: Four simple ways to apply behavioural insights*. The Behavioural Insights Team.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4–16. https://doi.org/10.3102/0013189X013006004

Cozman, F. G., & Kaufman, D. (2022). Viés no aprendizado de máquina em sistemas de inteligência artificial: A diversidade de origens e os caminhos de mitigação. *Revista USP*(135), 195–210. https://doi.org/10.11606/issn.2448-046X.v135i135p195-210

Guha, R. V., McCool, R., & Miller, E. (2003). Semantic search. In *Proceedings of the 12th International Conference on World Wide Web* (pp. 700–709). ACM. https://doi.org/10.1145/775152.775250

Huang, L., et al. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 1*(1), 1–58. https://doi.org/10.1145/XXXXXXX

Jacobs, A. Z., & Wallach, H. M. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 375–385). ACM. https://doi.org/10.1145/3442188.3445918

Kahneman, D. (2011). *Rápido e devagar: Duas formas de pensar*. Objetiva.

Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS), 34* (pp. 9459–9474). Curran Associates.

Melo, J. A. L. de. (2025). Racialização, viés algorítmico e poder na era da inteligência artificial: Uma análise crítica das novas arquiteturas tecnológicas. *Revista ARACÊ, 7*(12), 1–12.

Mohoney, J., et al. (2023). High-throughput vector similarity search in knowledge graphs. *Proceedings of the ACM on Management of Data, 1*(2), 1–25. https://doi.org/10.1145/XXXXXXX

Simões-Gomes, L., Roberto, E., & Mendonça, J. (2020). Viés algorítmico: Um balanço provisório. *Estudos de Sociologia, 25*(48), 139–166. https://doi.org/10.1590/eso.v25i48.20

Sunstein, C. R. (2021). *Sludge: What stops us from getting things done and what to do about it*. The MIT Press.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness (FairWare)* (pp. 1–7).
ACM. https://doi.org/10.1145/XXXXXXX