

Plataforma Inteligente para Ensino de Cibersegurança com Geração de Conteúdo Dinâmico

Matheus B. Honório, Herick José, Clarissa Santos, Danielle Rousy Dias Ricarte, Raoni Kulesza

Universidade Federal da Paraíba

João Pessoa - PB, Brasil

{matheus.honorio,clarissa.lima,raoni}@lavid.ufpb.br,danielle@ci.ufpb.br,hjf@academico.ufpb.br

ABSTRACT

Cybersecurity education requires dynamic and continuously updated learning materials, given the rapid obsolescence of traditional resources. This work presents a Retrieval-Augmented Generation (RAG) architecture integrated into the <omitted> educational platform to enable the dynamic generation of cybersecurity content. The solution combines semantic retrieval, prompt engineering, and automated evaluation through an LLM-as-a-Judge pipeline to produce multimodal content—including explanations, quizzes, flashcards and slides—derived from a specialized knowledge base of more than 2,100 pages. The approach was assessed across realistic instructional scenarios, with attention to groundedness, retrieval precision, pedagogical quality, and response adequacy. Results show 92% precision, 82% groundedness, and an 88.89% approval rate according to the evaluation rubric, indicating that the system generates accurate, contextualized, and educationally aligned content. The proposed architecture reduces material obsolescence, enhances personalization, and demonstrates the potential of RAG systems to support learning in complex domains such as cybersecurity.

KEYWORDS

Cybersecurity Education, Dynamic Content Generation, Gamification, Intelligent Tutoring Systems, Large Language Models, Retrieval-Augmented Generation

1 INTRODUÇÃO

A crescente digitalização no Brasil — onde mais de 93% dos domicílios têm acesso à internet — amplia a superfície de ataque cibernético e torna a literacia em segurança digital uma necessidade transversal [7]. Entretanto, o ensino tradicional de cibersegurança enfrenta um obstáculo estrutural: a rápida obsolescência dos materiais didáticos diante da evolução contínua de vulnerabilidades, técnicas de ataque e ferramentas de defesa. Abordagens estáticas falham em preparar os estudantes para esse cenário dinâmico e contribuem para a desmotivação e a baixa taxa de retenção [16].

Modelos de Linguagem de Grande Escala (LLMs) surgem como alternativa promissora para atualização de conteúdo,

mas sua adoção isolada em domínios críticos acarreta riscos severos de alucinações técnicas, como comandos inexistentes, parâmetros incorretos ou práticas inseguras [16]. Assim, persiste uma lacuna na literatura e no mercado em relação a ecossistemas educacionais capazes de unir precisão factual, atualização contínua e metodologias que promovam engajamento.

Este trabalho investiga como gerar conteúdos educacionais de cibersegurança que sejam simultaneamente atualizados, tecnicamente corretos e adaptados ao contexto do estudante. Esse desafio é complexo devido à volatilidade do domínio e às limitações dos modelos generativos.

Para enfrentar essa lacuna, propomos o ViTuRi, uma plataforma inteligente gamificada baseada em videobot, fundamentada em uma arquitetura de *Retrieval-Augmented Generation* (RAG). A abordagem desacopla a base factual dos pesos do modelo, garantindo precisão, rastreabilidade e geração dinâmica de múltiplos formatos educacionais, como textos, slides, flashcards e quizzes.

Este artigo apresenta quatro contribuições principais: (i) uma arquitetura RAG multimodal para ensino de cibersegurança, integrada a uma base de conhecimento especializada com mais de 2.000 páginas; (ii) um microserviço independente para recuperação vetorial e geração dinâmica de conteúdo educacional; (iii) um pipeline de avaliação baseado em *LLM-as-a-Judge*, com oito critérios de qualidade técnica e didática, e por fim, (iv) uma análise quantitativa da recuperação semântica e da qualidade das respostas, incluindo métricas de Precision, Recall, F1-Score e groundedness.

O restante do artigo aborda fundamentos teóricos (Seção 2), trabalhos relacionados (Seção 3), metodologia (Seção 4), resultados (Seção 5), limitações (Seção 6) e conclusões (Seção 7).

2 FUNDAMENTAÇÃO TEÓRICA E CONCEITUAL

A construção de um ecossistema inteligente voltado à formação em segurança digital exige a convergência de conceitos fundamentais de proteção de dados, metodologias pedagógicas adaptativas e engenharia de inteligência artificial de ponta. Esta seção delimita os marcos teóricos que sustentam a arquitetura da plataforma proposta.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.
ISSN 2966-2753

2.1 O Dinamismo na Segurança da Informação e os Desafios Educacionais

A Segurança da Informação (SI) é classicamente regida pela tríade CID — Confidencialidade, Integridade e Disponibilidade —, expandindo-se posteriormente para englobar a autenticidade e o não-repúdio (NIST 2020) [17]. No entanto, o domínio técnico da SI caracteriza-se por uma assimetria informacional volátil: novas vulnerabilidades (*Common Vulnerabilities and Exposures* - CVE) [14] e táticas de ataque (mapeadas por frameworks como o MITRE ATT&CK)[15] emergem diariamente.

No cenário educacional, o ensino de cibersegurança esbarra no gargalo dos currículos estáticos e manuais de laboratório tradicionais. Conforme apontado por Mukherjee et al. [16], abordagens pedagógicas passivas falham em preparar o estudante para o panorama real de ameaças, uma vez que o conteúdo programático se torna obsoleto rapidamente, reforçando a necessidade de Sistemas Tutores Inteligentes (ITS) capazes de gerar cenários atualizados e personalizados.

2.2 Modelos de Linguagem de Grande Escala (LLMs) e o Fenômeno da Alucinação

Paralelamente, os avanços recentes em Processamento de Linguagem Natural (PLN), impulsionados pela arquitetura *Transformer* e seus mecanismos de atenção [24], permitiram o surgimento dos *Large Language Models* (LLMs). Modelos como o GPT-4o mini demonstram capacidade de interpretar instruções complexas, sintetizar explicações e gerar código, tornando-se ferramentas promissoras de apoio educacional.

Entretanto, quando aplicados a domínios de alta severidade técnica — como a cibersegurança — os LLMs enfrentam limitações estruturais que comprometem sua confiabilidade:

- **Desatualização temporal:** O corte de conhecimento (knowledge cutoff) impede que o modelo nativo conheça vulnerabilidades de dia zero (zero-day) descobertas após o término do seu treinamento.
- **Alucinações estocásticas:** ao otimizar a probabilidade textual, e não a veracidade factual, o modelo pode inventar comandos, parâmetros ou práticas inseguras, como alertado por Mukherjee et al. [16].
- **Ausência de rastreabilidade:** as respostas geradas não indicam claramente a origem das informações, o que dificulta a validação pedagógica.

Essas limitações tornam arriscada a utilização de LLMs isolados em ambientes educacionais de cibersegurança, nos quais a precisão técnica e a atualização contínua são requisitos essenciais. Assim, torna-se necessário adotar arquiteturas que controlem o contexto factual e reduzam a dependência do conhecimento interno do modelo.

2.3 Arquitetura Retrieval-Augmented Generation (RAG)

A arquitetura *Retrieval-Augmented Generation* (RAG), formalizada por Lewis et al. [11], surge como solução para mitigar alucinações e garantir alinhamento factual. O RAG

desacopla o conhecimento do domínio dos pesos internos do modelo gerador, estruturando o pipeline em duas etapas complementares:

- **Recuperação (*retrieval*):** o sistema intercepta a consulta do estudante e busca, em um repositório externo confiável, os documentos mais relevantes — como guias técnicos, normas e materiais pedagógicos.
- **Geração (*generation*)** os fragmentos recuperados são injetados explicitamente no prompt, orientando o LLM a produzir respostas fundamentadas e rastreáveis.

Ao substituir o conhecimento interno do modelo por evidências extraídas da base factual, o RAG reduz drasticamente o risco de alucinações e permite atualização contínua sem necessidade de retreinamento. Uddin e Sarker [23] destacam que, no contexto da modelagem de cibersegurança, o RAG é superior ao *fine-tuning* por eliminar custos computacionais, preservar a rastreabilidade e garantir a aderência ao domínio.

Além disso, o RAG favorece a personalização pedagógica, pois o contexto recuperado pode ser ajustado conforme o nível de proficiência do aluno, o tipo de material solicitado (texto, slides, flashcards, quiz) e o objetivo instrucional. Essa flexibilidade torna o RAG especialmente adequado para ambientes educacionais dinâmicos, nos quais precisão, adaptabilidade e transparência são fundamentais.

2.4 Representação Vetorial

A etapa de recuperação em arquiteturas RAG depende da capacidade de representar documentos textuais em espaços vetoriais de alta dimensionalidade. Modelos de *embeddings* mapeiam trechos de texto em vetores que capturam relações semânticas, permitindo identificar similaridade conceitual mesmo quando não há correspondência lexical direta [11]. Essa abordagem supera mecanismos tradicionais baseados em palavras-chave, reduzindo o impacto de sinônimos, variações terminológicas e diferenças de estilo, fenômenos frequentes em domínios técnicos como a cibersegurança [2].

A busca por vizinhos mais próximos (*nearest-neighbor search*), geralmente realizada com métricas como a similaridade de cosseno, permite recuperar fragmentos altamente relevantes para a consulta do estudante [28]. Estudos recentes demonstram que *embeddings* contextuais elevam significativamente a precisão de sistemas de recuperação e recomendação em ambientes multimídia [18], tornando-se um componente essencial para pipelines RAG.

No contexto deste trabalho, a representação vetorial é empregada para garantir que o conteúdo recuperado seja semanticamente alinhado ao tópico solicitado, servindo de base factual para a geração de materiais educacionais [16]. Essa etapa é particularmente importante na cibersegurança, em que conceitos podem ser expressos por siglas, termos técnicos equivalentes ou descrições procedurais, o que exige mecanismos robustos de correspondência semântica [23].

2.5 RAG como Motor Cognitivo para Agentes Pedagógicos Multimídia

O desacoplamento entre recuperação e geração permite que o RAG atue como motor cognitivo em ecossistemas educacionais multimídia — videobots, chatbots instrucionais, ambientes gamificados — respondendo de forma dinâmica e alinhada à interação do aluno [10]. O RAG processa a ação do estudante, recupera evidências na base factual e sintetiza respostas fundamentadas, garantindo precisão técnica, rastreabilidade e adaptação pedagógica [8]; a camada de IA opera como microserviço independente, com alta disponibilidade e baixa latência, enquanto o ambiente educacional foca na experiência do usuário e na gamificação. Assim, o RAG complementa o ecossistema — sem substituir os fluxos tradicionais de autoria — fornecendo conteúdo atualizado e adaptado ao nível do estudante, reforçando princípios andragógicos como autonomia e aprendizagem orientada à prática.

3 TRABALHOS RELACIONADOS

Os trabalhos relacionados ao ensino de cibersegurança com suporte de IA podem ser organizados em três eixos complementares: (i) personalização e recomendação de conteúdo; (ii) sistemas tutores inteligentes e ambientes adaptativos; e (iii) abordagens baseadas em RAG aplicadas à educação em cibersegurança.

3.1 Personalização e Recomendação de Conteúdo

A personalização tem sido um tema recorrente no WebMedia, com trabalhos que exploram desde a distribuição multimídia até a gamificação narrativa. Araújo [1] introduziu um modelo pioneiro de personalização de conteúdo em ambientes educacionais ubíquos, utilizando perfis de usuário para adaptar materiais multimídia. Toledo Palomino et al. [22] avançaram essa linha ao propor um framework de gamificação narrativa que integra ontologias e estratégias de personalização, destacando a importância de adaptar não apenas o ambiente, mas também o próprio conteúdo instrucional.

Com a popularização dos LLMs, novas abordagens surgiram para a recuperação semântica e a interação natural. Campos et al. [2] demonstraram o potencial de LLMs para responder a perguntas sobre bases heterogêneas, enquanto e Silva et al. [4] apresentaram o EurekaChat, um chatbot multiagente que adapta suas respostas ao perfil do usuário, mantendo a conformidade com os requisitos de privacidade. Esses trabalhos reforçam a relevância da personalização, embora não abordem diretamente a geração de materiais educacionais estruturados.

3.2 Sistemas Tutores Inteligentes e Ambientes Adaptativos

A literatura sobre Sistemas Tutores Inteligentes (STI) apresenta uma evolução contínua de arquiteturas adaptativas. Cannataro e Pugliese [3] propuseram uma arquitetura de hipermídia baseada em grafos ponderados e em modelos probabilísticos de usuários, estabelecendo fundamentos para

sistemas que ajustam conteúdo conforme o comportamento do aprendiz. No contexto brasileiro, o AdaptWeb® [6] explorou hierarquias conceituais e pré-requisitos para a adaptação de conteúdo, contribuindo para a catalogação e reutilização de objetos de aprendizagem em múltiplas plataformas.

Esses trabalhos evidenciam a preocupação histórica da comunidade com mecanismos de adaptação, mas não exploraram o uso de modelos generativos ou de pipelines RAG para atualização dinâmica de materiais.

3.3 Abordagens RAG para Educação e Ensino de Cibersegurança

O uso de RAG na educação em cibersegurança é recente, mas já apresenta avanços relevantes. Zhao et al. [26] introduziram o CyberBOT, um chatbot que utiliza RAG combinado com ontologias para validar respostas e reduzir alucinações, demonstrando eficácia em cursos de pós-graduação. O framework CyberRAG [25] propõe uma abordagem em duas etapas — recuperação e validação ontológica — para mitigar incertezas em ambientes de aprendizagem baseada em problemas.

O RAG-PRISM [21] explora tutoria adaptativa com IA generativa, utilizando métricas, como o MRR, para personalizar o treinamento em cibersegurança. Os resultados preliminares mostram um elevado alinhamento entre as respostas geradas e os materiais instrucionais.

Embora promissores, esses trabalhos concentram-se em chatbots de perguntas e respostas e não abordam a geração de múltiplos formatos educacionais nem ciclos de feedback contínuo.

3.4 Lacunas e Contribuição do Trabalho

Apesar dos avanços, persistem lacunas importantes: (i) ausência de sistemas capazes de gerar múltiplos formatos de conteúdo (texto, slides, flashcards, quizzes, metadados de vídeo) a partir de uma única base vetorial; (ii) falta de pipelines que integrem feedback explícito e implícito dos alunos para retroalimentação contínua; e (iii) pouca exploração de RAG em ecossistemas multimídia com videobots e gamificação.

O presente trabalho preenche essas lacunas ao propor uma plataforma inteligente baseada em RAG que gera dinamicamente oito tipos de conteúdo educacional, incorpora sinais de qualidade provenientes das interações dos alunos e utiliza métricas como *groundedness*, diversidade controlada e latência P95 para monitoramento contínuo. Essa abordagem amplia o estado da arte ao integrar recuperação semântica, geração multimodal e adaptação pedagógica em um único ecossistema escalável.

Além disso, a própria versão anterior da plataforma Vi-TuRi, baseada em conteúdo estático gerado manualmente ou por LLMs com prompts fixos, representa um baseline relevante que evidencia a necessidade de uma abordagem dinâmica e fundamentada como a proposta neste trabalho.

4 METODOLOGIA

Esta pesquisa possui natureza aplicada e caráter exploratório, investigando o uso de RAG na geração automática de conteúdos educacionais para o ensino de cibersegurança. O objetivo é avaliar a capacidade de uma arquitetura baseada em recuperação semântica e modelos de linguagem de grande porte em produzir materiais educacionais contextualizados e aderentes a uma base de conhecimento especializada.

4.1 Arquitetura da solução

A solução foi implementada como um microserviço independente denominado *RAG Content Engine* (Figura 1), desenvolvido em Python 3.11¹ utilizando o framework FastAPI e integrado ao backend do ViTuRi por meio de uma API REST. A arquitetura é composta por três etapas principais: (i) construção da base de conhecimento; (ii) recuperação semântica de documentos relevantes; e (iii) geração de conteúdo educacional por meio de modelos de linguagem. Durante a geração, o contexto recuperado é enriquecido com informações provenientes de interações anteriores dos usuários, permitindo incorporar sinais de qualidade ao processo de produção de conteúdo.

4.2 Construção da base de conhecimento

A base de conhecimento foi composta por 30 documentos do Programa Hackers do Bem², totalizando 2.152 páginas e aproximadamente 449 mil palavras, abrangendo temas como riscos e compliance (GRC), criptografia, redes, computação em nuvem, desenvolvimento seguro, inteligência de ameaças e resposta a incidentes.

Após a extração textual, os documentos foram segmentados com o algoritmo *RecursiveCharacterTextSplitter*³ (400 caracteres, sobreposição de 100), equilibrando a granularidade e a preservação de contexto. Cada segmento foi convertido em um vetor de 1.536 dimensões utilizando o modelo *text-embedding-3-small*⁴ da OpenAI, e armazenado no MongoDB Atlas Vector Search, configurado com similaridade de cosseno.

Essa etapa permitiu construir um repositório vetorial robusto, adequado para consultas semânticas em um domínio técnico com vocabulário denso e terminologia variável.

4.3 Recuperação Semântica de Documentos Relevantes

Para cada requisição — pergunta, solicitação de resumo ou geração de quiz —, a consulta é convertida em um *embedding* e comparada ao índice vetorial por meio da similaridade de cosseno. O sistema retorna os k fragmentos mais relevantes, ajustando dinamicamente o valor de k conforme: (i) tipo

de conteúdo solicitado (texto, slides, flashcards, quiz); (ii) nível de proficiência do estudante (iniciante, intermediário, avançado). Esse mecanismo permite que o sistema recupere informações contextualmente alinhadas, mesmo quando os termos exatos não estão presentes nos documentos — característica especialmente relevante no domínio da cibersegurança, onde conceitos podem ser expressos por diferentes terminologias, siglas ou sinônimos (por exemplo, firewall, filtro de pacotes, ACL).

Metadados, como o título da seção e as palavras-chave, são incorporados ao contexto, enriquecendo a representação semântica e aumentando a precisão da recuperação — o que é especialmente importante na cibersegurança, onde sinônimos e siglas são frequentes.

4.4 Geração de Conteúdo Educacional

A geração é conduzida pelo modelo GPT-4o-mini [19], que recebe: (i) a consulta original, (ii) os fragmentos recuperados e, (iii) instruções específicas para cada formato. A arquitetura suporta qualquer LLM compatível com a interface OpenAI, incluindo modelos open-source como Llama e Qwen.

O sistema produz oito tipos de materiais educacionais: texto dissertativo, slides, flashcards, quizzes, metadados para vídeos, referências para PDFs e conteúdo introdutório composto. Cada formato utiliza uma engenharia de prompt própria, ajustada ao nível de proficiência do estudante.

Para garantir aderência factual, o modelo é instruído a priorizar as informações dos fragmentos recuperados e a sinalizar explicitamente quando recorrer ao conhecimento interno. Esse mecanismo reduz alucinações técnicas, como comandos inexistentes ou práticas inseguras.

4.5 Perfil de Usuário e Interação

O ViTuRi é direcionado a estudantes do Programa Hackers do Bem — de jovens a partir de 16 anos a profissionais em transição para a cibersegurança —, atuando como ferramenta complementar de microaprendizagem: unidades curtas de conteúdo, consumidas sob demanda, que favorecem a retenção sem exigir uma nova sessão extensa de estudo. A interação ocorre por meio de componentes específicos para cada formato: no flashcard, a frente exibe o tópico e o verso revela a resposta ao ser tocado; no quiz, o estudante seleciona uma alternativa e recebe feedback textual imediato de acerto ou erro.

4.6 Protocolo de Avaliação

A avaliação seguiu uma matriz combinatória entre os seis tipos de conteúdo gerados e os três níveis de proficiência do estudante, totalizando 18 cenários de teste; vídeo e externo foram excluídos por o sistema gerar apenas conteúdo textual, sem recursos multimídia ou links externos na base indexada (Seção 4.2). Cada cenário combina uma camada determinística — código que verifica falhas estruturais que inviabilizariam o uso em produção — e uma camada semântica, julgada pelo modelo árbitro via *rubric prompting* específico

¹Python 3.11. Disponível em: <https://www.python.org/>

²Detalhamento da base de conhecimento (documentos, estrutura por nível/tópico e metadados de indexação). Disponível em: <https://mthonorio.github.io/vituri-knowledge-base/>

³*RecursiveCharacterTextSplitter*. Disponível em: <https://reference.langchain.com/python/langchain-text-splitters/character/RecursiveCharacterTextSplitter>

⁴*text-embedding-3-small*. Disponível em: <https://developers.openai.com/api/docs/models/text-embedding-3-small>

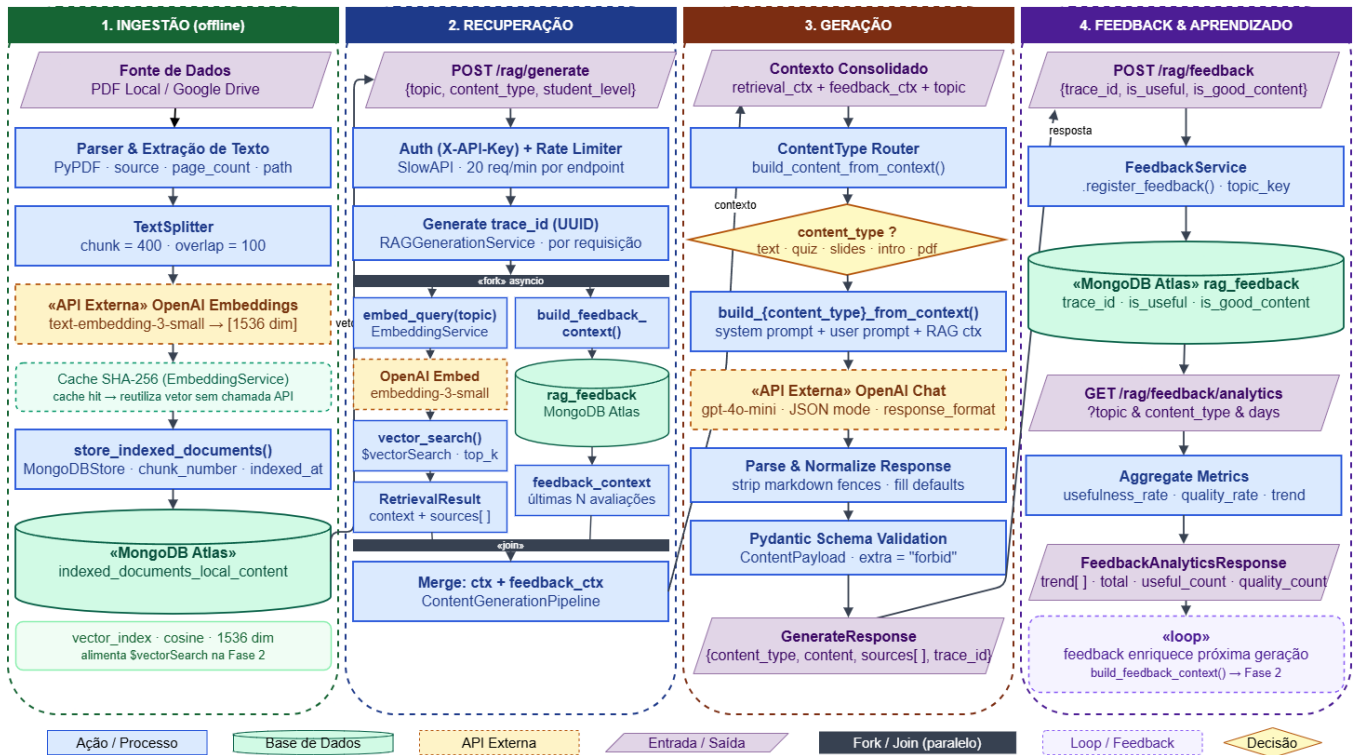


Figure 1: Pipeline RAG (4 fases).

ao tópico e ao tipo de conteúdo. Os critérios de julgamento e o cálculo das métricas são detalhados na Seção 5.1.

4.7 Baseline de Comparação

Antes da adoção da arquitetura RAG, a plataforma ViTuRi operava com um modelo tradicional de geração de conteúdo educacional. Nessa versão, os materiais eram produzidos manualmente por professores ou gerados por LLMs com prompts fixos e altamente específicos. Uma vez criado, o conteúdo permanecia estático, sendo reutilizado de forma idêntica para todos os estudantes do mesmo nível de proficiência, independentemente do contexto da interação ou da evolução do domínio da cibersegurança.

Essa abordagem apresentava limitações estruturais: ausência de atualização contínua, dependência de autoria manual, impossibilidade de personalização, risco de inconsistências entre módulos e falta de rastreabilidade das fontes. Além disso, o uso de LLMs sem recuperação semântica aumentava a probabilidade de alucinações técnicas, especialmente em tópicos avançados.

Por representar o estado anterior da plataforma e refletir práticas comuns em ambientes educacionais digitais, essa versão é utilizada como baseline para comparação com a solução RAG proposta neste trabalho. A Tabela 1 resume as principais diferenças entre as duas abordagens.

Table 1: Baseline comparativo.

Item	Baseline	RAG Proposto
Geração	Manual ou LLM fixo	Automática (RAG+LLM)
Atualização	Estática	Contínua (corpus)
Personalização	Por nível	Nível + Contexto
Formatos	8 modos	8 modos
Rastreabilidade	Não	Evidências
Escalabilidade	Baixa	Alta
Alucinação	Moderado	Mitigado
Dependência de professores	Alta	Reduzida

5 ANÁLISE DE TESTES E RESULTADOS

Esta seção apresenta os resultados das avaliações realizadas sobre o sistema ViTuRi RAG, organizados em três dimensões complementares: (i) qualidade das respostas geradas, avaliada por meio de uma infraestrutura *LLM-as-a-Judge*; (ii) indicadores-chave de desempenho (KPIs) relacionados à experiência do estudante e à aderência pedagógica; e (iii) métricas de recuperação vetorial que mensuram a precisão técnica da etapa de retrieval. Ao final, uma comparação com o baseline tradicional da plataforma ViTuRi evidencia os ganhos proporcionados pela arquitetura RAG.

5.1 Avaliação com LLM-as-a-Judge

A qualidade das respostas foi avaliada por meio da estratégia *LLM-as-a-Judge* [27], preferida a métricas lexicais (BLEU, ROUGE, METEOR), que não capturam correção factual nem utilidade pedagógica [12], e a *frameworks* específicos para RAG como o RAGAS [5], que não cobrem conformidade estrutural nem adequação pedagógica multiformato. Optou-se por uma rubrica própria, implementada com PromptFoo [20] e o modelo Claude Haiku 4.5⁵ como árbitro (Figura 2), com oito critérios (C1–C8) em dois níveis de peso: **ALTO** (C1, C2, C3, C7) e **MÉDIO** (C4, C5, C6, C8), totalizando 20 pontos — separação que isola falhas estruturais determinísticas das dimensões semânticas julgadas pelo modelo, mitigando o viés de juízes LLM de favorecer respostas mais longas [27]. A aprovação exige conformidade em todos os critérios de peso ALTO (Tabela 2)

Table 2: Critérios do LLM-as-a-Judge (rubrica C1–C8).

ID	Critério	Peso	Pontos
C1	Conformidade com o esquema esperado da solução	ALTO	3
C2	Campo <i>content_type</i> corresponde à requisição	ALTO	3
C3	Relevância temática ao tópico solicitado	ALTO	3
C4	Qualidade educacional, correção factual e utilidade	MÉDIO	2
C5	Presença e validade do vetor de fontes (<i>sources</i>)	MÉDIO	2
C6	Clareza e consistência linguística em PT-BR	MÉDIO	2
C7	Integridade do envelope de resposta e UUID válido	ALTO	3
C8	Compleitude da cobertura do tópico	MÉDIO	2
Total			20

Os 18 cenários da suíte de testes (Seção 4.6) cobriram tópicos centrados em domínios de cibersegurança como criptografia, vulnerabilidades web e arquitetura Zero Trust.

O sistema obteve **88,89%** de aprovação (16 de 18 cenários), demonstrando alta conformidade estrutural e relevância temática, especialmente nos critérios de maior peso.

5.2 Indicadores-Chave de Desempenho (KPIs)

Paralelamente à avaliação por juiz, mensuraram-se os cinco indicadores de desempenho definidos para o projeto, apresentados na Tabela 3.

A única não conformidade foi a *latência P95*, que atingiu 14s (meta ≤ 10 s). O profiling indica que a busca vetorial em 1.536 dimensões responde por 40% do tempo total, seguida pela geração de *embeddings* e pela chamada ao modelo gerador. Apesar disso, todos os demais KPIs foram plenamente atendidos.

⁵Claude Haiku 4.5. Disponível em: <https://www.anthropic.com/claude/haiku>

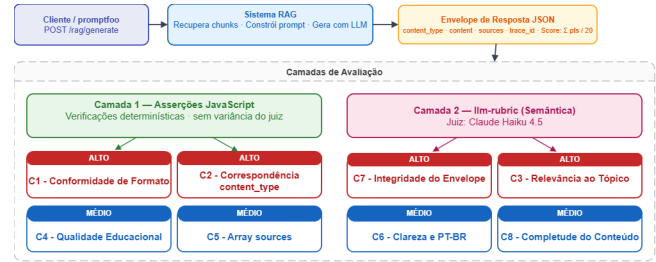


Figure 2: Pipeline de avaliação automatizada LLM-as-a-Judge no módulo RAG. Fluxo desde a requisição ao endpoint POST /rag/generate até o julgamento pelos oito critérios de aceite (C1–C8) via Claude Haiku 4.5.

Table 3: Resultado dos KPIs.

KPI	Meta	Resultado	Status
Relevância percebida (média)	$\geq 4,05$	5,005	✓
Taxa de aprovação (feedback)	$\geq 70\%$	88%	✓
<i>Groundedness</i>	$\geq 80\%$	82%	✓
Diversidade controlada	5%–80%	60%	✓
Latência P95	≤ 10 s	14 s	×

5.3 Métricas de Recuperação Vetorial

A qualidade da etapa de recuperação (*retrieval*) foi mensurada por meio de um protocolo de avaliação desenvolvido especificamente para este trabalho, de forma independente da etapa de geração. A metodologia adota duas métricas clássicas de recuperação de informação adaptadas ao contexto RAG:

$$\text{Precision} = \frac{|\text{recuperados relevantes}|}{|\text{recuperados}|} \quad (1)$$

$$\text{Recall} = \frac{|\text{recuperados relevantes}|}{|\text{relevantes no pool}|} \quad (2)$$

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Como o universo de documentos relevantes para cada consulta é desconhecido *a priori*, o denominador do Recall foi aproximado por *pool* de $k_{\text{pool}} = 50$ documentos recuperados por similaridade de cosseno, tratados como o conjunto máximo de candidatos relevantes. Essa aproximação subestima o universo real de documentos relevantes presentes no corpus, de modo que o Recall reportado deve ser interpretado como um limite inferior do valor real — o que não afeta a validade comparativa entre configurações, mas impede generalizações absolutas sobre cobertura. A relevância foi determinada pelo critério de limiar de score: documentos com pontuação de similaridade $\geq 0,75$ foram classificados como relevantes (*modo score*, sem custo adicional de LLM). A avaliação utilizou $k_t = 5$ documentos para o cálculo da Precision, sobre 10 consultas em PT-BR nos domínios de cibersegurança presentes no corpus — uma amostra piloto que será ampliada

em trabalhos futuros para fortalecer a robustez estatística das métricas.

Table 4: Métricas de recuperação vetorial.

Configuração	Precision	Recall	F1-Score
$k_t = 5$, $k_{pool} = 50$, $th = 0,75$	92,0%	36,3%	52,1%

Os resultados evidenciam uma **Precision elevada** (92,0%): a grande maioria dos documentos efetivamente recuperados dentro do top-5 é genuinamente relevante ao tópico consultado. Contudo, o **Recall reduzido** (36,3%) indica que apenas uma fração dos documentos potencialmente úteis presentes no corpus é capturada em cada consulta. Isso caracteriza um fenômeno de *vocabulary mismatch*: a formulação exata da consulta em linguagem natural pode não corresponder às variações lexicais com que o mesmo conceito é expresso nos documentos indexados (e.g., “phishing” versus “ataque de engenharia social” [13]).

O F1-Score de 52,1% reflete esse desequilíbrio: embora a recuperação seja precisa, ela é *incompleta*. A principal causa estrutural identificada é o valor padrão $k_t = 3$ adotado na pipeline de geração (`max_sources`), que limita o teto aritmético do Recall a aproximadamente 20–40% mesmo com busca perfeita. Como direção de melhoria, propõe-se a adoção de *Multi-Query Retrieval* [9], técnica que expande a consulta original em N variantes por meio de um LLM e mescla os resultados, mitigando o *vocabulary mismatch* e elevando potencialmente o Recall para a faixa de 55–65%.

5.4 Comparação com o Baseline Tradicional

Para avaliar os ganhos proporcionados pela arquitetura RAG, os resultados obtidos foram comparados aos da versão anterior da plataforma, que utilizava conteúdo estático gerado manualmente ou por LLMs com prompts fixos. A comparação considera três dimensões: qualidade das respostas, desempenho técnico e escalabilidade do sistema.

5.4.1 Qualidade das Respostas. A versão tradicional apresentava variação significativa na precisão técnica, dependência da habilidade do professor em formular *prompts* eficazes e ausência de mecanismos de *groundedness*. Em contraste, a versão com RAG alcançou resultados superiores em todas as métricas avaliadas, conforme detalhado na Tabela 5.

Table 5: Comparação de qualidade entre Baseline e RAG

Métrica	Baseline	RAG
Precisão técnica	Variável	92%
Groundedness	Não mensurado	82%
Conformidade estrutural	Inconsistente	Alta
Risco de alucinação	Moderado	Baixo
Diversidade controlada	Não aplicável	60%

Para validar esses ganhos em um cenário específico, foi conduzido um teste comparativo com o tópico *Criptografia*

simétrica e o algoritmo *RC4 (PPGA)*, utilizando o mesmo modelo gerador (GPT-4o-mini) e juiz avaliador (Claude Haiku 4.5). Os resultados, apresentados na Tabela 6, confirmam a superioridade da arquitetura RAG mesmo em tópicos menos comuns.

Table 6: Comparação específica para o tópico RC4 (quiz, nível iniciante)

Métrica	Baseline	RAG
Precisão técnica	4,0/5	5,0/5
Groundedness	85%	95%
Conformidade estrutural	Média	Média
Risco de alucinação	Baixo	Baixo
Diversidade controlada	64%	61%

O ganho mais expressivo foi na precisão técnica (4,0→5,0), evidenciando que o acesso à base especializada gera respostas mais corretas mesmo em tópicos menos recorrentes; o aumento do *groundedness* (85%→95%) confirma a eficácia da recuperação semântica na ancoragem factual.

5.4.2 Desempenho Técnico. A versão tradicional não possuía recuperação vetorial, o que impossibilitava a medição de métricas como *Precision*, *Recall* e F1-Score. O modelo RAG apresentou *Precision* de 92%, embora com *Recall* de 36,3%, indicando espaço para melhorias na cobertura semântica (Tabela 7).

Table 7: Comparação técnica entre Baseline e RAG

Métrica	Baseline	RAG
Recuperação vetorial	Não existe	Sim
<i>Precision</i>	Não aplicável	92%
<i>Recall</i>	Não aplicável	36,3%
F1-Score	Não aplicável	52,1%
Latência P95	10 s	14 s

Embora a latência P95 tenha aumentado de 10s para 14s, esse incremento é justificado pelo ganho em precisão e atualização de conteúdo trazido pela recuperação vetorial e geração dinâmica, restando como oportunidade de otimização futura (Seção 7).

5.4.3 Síntese Comparativa. A comparação evidencia que o ViTuRi RAG supera o baseline em praticamente todas as dimensões avaliadas: maior precisão técnica (92% contra variabilidade no baseline), maior aderência factual (*groundedness* de 95% contra 85%), menor risco de alucinações (evidências recuperadas), personalização real (nível do aluno e contexto), atualização contínua (independência do *knowledge cutoff*), rastreabilidade das fontes e diversidade de formatos (oito tipos de conteúdo gerados automaticamente).

Apesar desses avanços, dois desafios operacionais se destacam: (i) a latência elevada (P95 de 14s, acima dos 10s esperados), que compromete a fluidez das interações síncronas

com o ViTuRi; e (ii) a cobertura semântica restrita (*Recall* de 36,3%), decorrente do desalinhamento lexical entre consultas e documentos, o que limita a abrangência da base recuperada.

Os resultados confirmam que, mesmo com as limitações identificadas, a substituição do modelo estático baseado em *prompts* fixos pela arquitetura RAG representa um avanço substancial na qualidade, confiabilidade e adaptabilidade do conteúdo educacional, consolidando a plataforma como uma solução viável para o ensino dinâmico de cibersegurança.

6 LIMITAÇÕES

Apesar dos resultados positivos, o sistema apresenta limitações estruturais que apontam oportunidades de evolução. A principal delas é a latência da etapa de recuperação vetorial, evidenciada pelo KPI de Latência P95 (14 s), acima da meta de 10 s. A busca por similaridade de cosseno em vetores de 1.536 dimensões, sobre uma base de 2.152 páginas, impõe um custo computacional significativo, o que impacta diretamente a fluidez das interações síncronas com o videobot. Esse gargalo é crítico em ambientes multimídia, nos quais a experiência pedagógica depende de respostas rápidas e contínuas.

Outra limitação é a ausência de síntese automática de vídeo: embora o sistema gere metadados estruturados (roteiros, legendas, instruções narrativas), a plataforma ViTuRi não integra modelos de Text-to-Video (T2V), o que exigiria acoplar ao pipeline um modelo externo especializado, com desafios de infraestrutura, custo e coerência pedagógica das cenas geradas.

O sistema também depende de recuperação semântica baseada em consultas únicas, o que contribui para o Recall reduzido (36,3%). O fenômeno de vocabulary mismatch — comum em cibersegurança — limita a cobertura de documentos relevantes, mesmo quando a precisão é elevada. Essa limitação é estrutural e decorre do valor padrão de *max_sources*, que limita o teto aritmético do Recall.

Por fim, a avaliação de qualidade apoia-se exclusivamente na infraestrutura *LLM-as-a-Judge* (Claude Haiku 4.5), sem validação cruzada por avaliadores humanos. Embora a rubrica de oito critérios ofereça objetividade e escalabilidade, a ausência de concordância inter-avaliadores (por exemplo, o Kappa de Cohen) entre o juiz automatizado e especialistas em cibersegurança ou educadores limita a validade externa dos resultados reportados quanto à qualidade pedagógica percebida.

7 CONCLUSÃO E TRABALHOS FUTUROS

Este artigo apresentou o ViTuRi, uma plataforma inteligente e gamificada para o ensino de cibersegurança, fundamentada em uma arquitetura de Retrieval-Augmented Generation (RAG). A solução endereça a rápida obsolescência dos materiais didáticos ao ancorar a geração de conteúdo em uma base de conhecimento especializada e atualizada, proveniente do Programa Hackers do Bem. A arquitetura demonstrou eficácia na produção de múltiplos formatos educacionais —

textos, slides, flashcards, quizzes e metadados para vídeos — com alta aderência técnica e factual.

Os resultados confirmam a robustez da abordagem: 88,89% de aprovação na avaliação LLM-as-a-Judge, 92,0% de Precisão, 82% de Groundedness, 5,00/5 de relevância percebida e 60% de diversidade controlada. Embora o Recall de 36,3% indique espaço para melhoria, a elevada precisão e aderência factual confirmam que o pipeline gera conteúdo confiável e contextualizado no domínio da cibersegurança.

A implementação do ciclo de feedback contínuo estabelece as bases para um sistema de aprendizagem adaptativo, reforçando o papel do RAG como motor cognitivo da plataforma e oferecendo uma infraestrutura escalável para ambientes educacionais multimídia.

Como direções de evolução, destacam-se:

- A adoção de Multi-Query Retrieval será priorizada para mitigar o vocabulary mismatch e elevar o Recall para 55–65%, com o conjunto de consultas de avaliação ampliado para além das 10 atuais.
- Para reduzir a Latência P95, serão exploradas estratégias como *caching* semântico, compressão de *prompts/embeddings* e busca híbrida (densa + lexical).
- Um estudo de validação humana com especialistas em cibersegurança e educadores medirá a concordância com o juiz automatizado (*LLM-as-a-Judge*) via Kappa de Cohen, complementando a avaliação automatizada.
- A substituição do GPT-4o-mini por modelos abertos como o Qwen-3.6⁶ permitirá reduzir custos e fortalecer a soberania da infraestrutura de IA.
- A incorporação de modelos T2V poderá viabilizar videoaulas personalizadas, e estudos com turmas-piloto medirão impacto pedagógico, engajamento e retenção em cenários reais.

Com essas evoluções, o ViTuRi avança rumo à construção de um ecossistema educacional em cibersegurança mais acessível, adaptativo e alinhado aos princípios da ciência aberta.

CONSIDERAÇÕES ÉTICAS

Este trabalho não envolveu seres humanos: a avaliação usou apenas métricas automatizadas (LLM-as-a-Judge e recuperação vetorial) sobre cenários sintéticos, sem coleta de dados pessoais. A base de conhecimento contém apenas materiais técnico-institucionais do Programa Hackers do Bem, sem dados pessoais ou sensíveis. O trabalho enquadra-se, portanto, nas hipóteses de não aplicabilidade da Resolução 510/2016 do CEP/CONEP, dispensando comitê de ética e CAAE. Os estudos com turmas-piloto propostos como trabalho futuro (Seção 7) serão submetidos a comitê de ética antes de sua execução.

DECLARAÇÃO SOBRE A UTILIZAÇÃO DE FERRAMENTAS E TECNOLOGIAS DE IA

O Claude (Anthropic) foi utilizado para revisão gramatical e ortográfica de uma versão preliminar deste manuscrito;

⁶Qwen-3.6. Disponível em: <https://qwen.ai>

tudo o conteúdo científico, interpretações e conclusões são de responsabilidade exclusiva dos autores.

AGRADECIMENTOS

O presente trabalho foi realizado com o apoio do RNP com recursos oriundos da Lei das TICs - Lei nº 8.248, de 23 de outubro de 1991.

REFERENCES

- [1] Rafael Dias Araújo. 2013. *Distribuição e personalização de conteúdo multimídia em ambientes educacionais ubíquos*. Dissertação de Mestrado. Universidade Federal de Uberlândia, Uberlândia. doi:10.14393/ufu.di.2013.309
- [2] João Campos, Marcelo Machado, Guilherme Lima, and Viviane Silva. 2025. Uma Ferramenta Web Baseada em LLMs para Respostas a Perguntas Simples sobre Bases de Conhecimento Heterogêneas. In *Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 159–162. doi:10.5753/webmedia_estendido.2025.16367
- [3] Mario Cannataro and Andrea Pugliese. 2000. An XML-based architecture for adaptive Web hypermedia systems using a probabilistic user model. In *Proceedings of the 1st International Conference on Web Information Systems Engineering*. IEEE, 1–8. doi:10.1109/WISE.2000.882386
- [4] Matheus Figueiredo e Silva, Levi Pereira Júnior, David Pereira, Beatriz Meneses, Kemilli Lima, José Araujo Junior, Leandro Marinho, and Eliane Araújo. 2025. EurekaChat: A Multi-Agent LLM-Based Chatbot for Supporting Students, Faculty, and University Administration. In *Anais Estendidos do XXXI Simpósio Brasileiro de Sistemas Multimídia e Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 107–110. doi:10.5753/webmedia_estendido.2025.16419
- [5] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217 [cs.CL] <https://arxiv.org/abs/2309.15217>
- [6] Isabela Gasparini, José Palazzo Moreira de Oliveira, Marcelo S. Pimenta, José Valdeni de Lima, Avani de Kemczinski, Mario L. Proença Jr, and Maria Angélica de O. Camargo Brunetto. 2010. AdaptWeb® - Evolução e Desafios. *Cadernos de Informática* 4, 2 (July 2010), 47–56. <https://seer.ufrgs.br/index.php/cadernosdeinformatica/article/view/v4n2p47-56>
- [7] IBGE. 2025. *Coordenação de Pesquisas por Amostra de Domicílios*. 2025. *Acesso à internet e à televisão e posse de telefone móvel celular para uso pessoal 2024*. Technical Report. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro. <https://biblioteca.ibge.gov.br>
- [8] Malcolm S. Knowles, Elwood F. Holton III, Richard A. Swanson, and Petra A. Robinson. 2020. *The Adult Learner: The Definitive Classic in Adult Education and Human Resource Development* (9th ed.). Routledge, New York, NY.
- [9] LangChain Contributors. 2024. MultiQueryRetriever. https://python.langchain.com/docs/how_to/MultiQueryRetriever.
- [10] James C. Lester, Sharolyn A. Converse, Susan E. Kahler, S. Todd Barlow, Brian A. Stone, and Ravinder S. Bhogal. 1997. The persona effect: affective impact of animated pedagogical agents. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) (CHI '97). Association for Computing Machinery, New York, NY, USA, 359–366. doi:10.1145/258549.258797
- [11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.
- [12] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. arXiv:2411.16594 [cs.AI] <https://arxiv.org/abs/2411.16594>
- [13] MITRE Corporation. 2023. MITRE ATT&CK: Phishing (T1566). <https://attack.mitre.org/techniques/T1566>.
- [14] MITRE Corporation. 2025. Common Vulnerabilities and Exposures (CVE®). <https://cve.mitre.org/>. Acesso em: 26 jun. 2026.
- [15] MITRE Corporation. 2025. MITRE ATT&CK®. <https://attack.mitre.org/>. Acesso em: 26 jun. 2026.
- [16] Madhav Mukherjee, John Le, and Yang-Wai Chow. 2025. Generative AI-Enhanced Intelligent Tutoring System for Graduate Cybersecurity Programs. *Future Internet* 17, 4 (2025), 154. doi:10.3390/fi17040154
- [17] National Institute of Standards and Technology. 2020. *Data Integrity: Detecting and Responding to Ransomware and Other Destructive Events*. Technical Report NIST SP 1800-26. National Institute of Standards and Technology. Disponível em: <https://doi.org/10.6028/NIST.SP.1800-26>. doi:10.6028/NIST.SP.1800-26
- [18] Francisco Ricarte Neto, Rafael Anchieta, Raimundo Moura, Pedro Santos Neto, and Andre Santana. 2025. Investigating Contextual Word Embeddings in Semi-Supervised Learning for Toxic Comment Detection. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 330–339. doi:10.5753/webmedia.2025.15994
- [19] OpenAI. 2024. GPT-4o mini: Advancing Cost-Efficient Intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>. Large language model. Accessed: 2026-06-25.
- [20] PromptFoo Contributors. 2024. PromptFoo: Test your prompts, agents, and RAGs. <https://promptfoo.dev>.
- [21] Gaurangi Raul, Yu-Zheng Lin, Karan Patel, Bono Po-Jen Shih, Matthew W. Redondo, Banafsheh Saber Latibari, Jesus Pacheco, Soheil Salehi, and Pratik Satam. 2025. RAG-PRISM: A Personalized, Rapid, and Immersive Skill Mastery Framework with Adaptive Retrieval-Augmented Tutoring. In *2025 IEEE Frontiers in Education Conference (FIE)*. 1–9. doi:10.1109/FIE63693.2025.11328742
- [22] Paula Toledo Palomino, Lennart Nacke, and Seiji Isotani. 2023. Gamificação Narrativa para Engajamento e Personalização: Redefinindo a Experiência do Aprendizado Digital. *Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*. doi:10.5753/webmedia_estendido.2023.233396
- [23] Mohammad Amaz Uddin and Iqbal H. Sarker. 2026. Fine-Tuning vs. RAG: A Position Paper from the Perspective of LLM-Based Cybersecurity Modeling. *Computing@AI Connect* 3 (2026), Article ID: 2026.0031. doi:10.69709/CAIC.2026.155073
- [24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
- [25] Chengshuai Zhao, Garima Agrawal, Fan Zhang, Tharindu Kumarage, Zhen Tan, Yuli Deng, Ying-Chih Chen, and Huan Liu. 2025. Ontology-Aware RAG for Improved Question-Answering in Cybersecurity Education. arXiv:2412.14191 [cs.CY] <https://arxiv.org/abs/2412.14191>
- [26] Chengshuai Zhao, Riccardo De Maria, Tharindu Kumarage, Kumar Satvik Chaudhary, Garima Agrawal, Yiwen Li, Jongchan Park, Ying-Chih Chen, Yuli Deng, and Huan Liu. 2025. CyberBOT: Ontology-Grounded Retrieval Augmented Generation for Reliable Cybersecurity Education. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*. ACM, 6752–6756. doi:10.1145/3746252.3761478
- [27] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) (NIPS '23). Curran Associates Inc., Red Hook, NY, USA, Article 2020, 29 pages.
- [28] Ítalo Pereira and Anderson Ferreira. 2024. E-BELA: Enhanced Embedding-Based Entity Linking Approach. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz

de Fora/MG). SBC, Porto Alegre, RS, Brasil, 115–123. doi:10.

5753/webmedia.2024.243160