S.N.A.K.E: Inteligência Artificial para Letramento Digital Ativo e Combate à Desinformação em Português

Lucas C. A. Melo[1], **Jean C. P. Cassiano**[1], **Robson Bonidia**[1,2], **André de Carvalho**[1]

[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
Av. Trabalhador São-carlense, 400, 13566-590, São Carlos-SP, Brasil

[2]Departamento de Computação – Universidade Tecnológica Federal do Paraná
Av. Alberto Carazzai, 1640 — Centro, Cornélio Procópio–PR, 86300-000, Brasil

Abstract. *Combating disinformation on messaging apps poses a dual challenge for the Brazilian population: the rapid propagation of fake news and users' limited digital literacy. While current solutions primarily focus on passive fact-checking, this work presents S.N.A.K.E., an integrated Artificial Intelligence-based virtual assistant designed with clear social purposes: promoting active digital literacy and fostering critical thinking. Unlike conventional approaches that focus solely on delivering a veracity verdict, S.N.A.K.E. serves as a digital tutor that, through interactive dialogs, explains the analysis criteria and subsequently empowers users to identify disinformation patterns. Preliminary results suggest that S.N.A.K.E. performs particularly well in explanation-oriented and pedagogical dimensions when compared with selected fact-checking tools.*

Resumo. *O combate à desinformação em aplicativos de mensagens enfrenta um desafio duplo para a população brasileira: a velocidade de propagação das fake news e a carência de letramento digital dos usuários. Enquanto as soluções atuais focam majoritariamente na verificação passiva de fatos (fact-checking), o presente trabalho apresenta o System for News Authentication and Knowledge Evaluation (S.N.A.K.E.), um assistente virtual integrado baseado em Inteligência Artificial, desenhado com propósitos sociais claros: promover o letramento digital ativo e fomentar o pensamento crítico. Diferentemente de abordagens convencionais que focam em classificadores, ou seja, em entregar apenas o veredito de veracidade, o S.N.A.K.E. atua como um tutor digital que, por meio de diálogos interativos, explica os critérios de análise e promove a capacitação do usuário no que tange à identificação de padrões de desinformação. Os resultados preliminares sugerem que o S.N.A.K.E. apresenta um desempenho particularmente robusto nas dimensões explicativa e pedagógica, quando comparado com ferramentas de verificação de fatos selecionadas.*

1. Introdução

A disseminação de desinformação em plataformas de mensagens e redes sociais apresenta-se como um desafio crítico para a integridade da informação. Como apontam Oliveira e Gomes [Oliveira and Gomes 2019], a propagação massiva de notícias falsas não apenas distorce o debate público, mas também desgasta os fundamentos da democracia ao explorar vulnerabilidades cognitivas e a falta de regulação eficaz nesses contextos.

Presentes no dia a dia da população, as plataformas de mensagens favorecem a circulação de notícias falsas em ambientes de confiança pessoal, onde a checagem de fatos é frequentemente negligenciada pelos usuários. Tais aplicativos (como WhatsApp e Telegram) operam como "espaços de mesocomunicação", onde a verificação de fatos fica mais difícil devido à privacidade e à confiança interpessoal entre os membros de grupos, tornando a correção de boatos uma tarefa complexa para as agências tradicionais [Frischlich et al. 2024].

Nesse contexto, soluções baseadas em Inteligência Artificial (IA) têm apresentado potencial para a detecção de notícias falsas e para o letramento digital. Além disso, a integração dessas ferramentas no processo educacional contra a desinformação tem um impacto positivo, atuando não apenas como suporte técnico, mas também como mediador que potencializa a aprendizagem colaborativa e crítica [Joseph and Thomas 2024].

Iniciativas pioneiras utilizam soluções de IA para combater esse fenômeno, tais como o *chatbot* "Fátima", desenvolvido pela agência Aos Fatos [Aos Fatos 2023], que exemplifica o uso de Grandes Modelos de Linguagem (LLMs, do inglês *Large Language Models*) para analisar e verificar informações, cruzando dados com checagens prévias da organização. Similarmente, a ferramenta "Tá Certo Isso Aí?" [Lira et al. 2025] utiliza IA para analisar notícias submetidas via WhatsApp, um aplicativo de mensagens comumente utilizado pelos brasileiros.

Apesar da robustez de tais ferramentas, elas operam predominantemente sob uma lógica classificatória a partir de um *input*: o usuário submete a dúvida e recebe um veredito de veracidade. Por outro lado, a literatura sugere que, para os chatbots serem aceitos e efetivos como contramedidas à desinformação, é necessário superar a desconfiança do usuário por meio da transparência e da educação, evitando que sejam percebidos unicamente como "caixas-pretas" que somente ditam a verdade [Frischlich et al. 2024].

A interação guiada com IA pode fomentar o pensamento crítico, desde que a ferramenta seja desenhada para desafiar o usuário cognitivamente e não apenas fornecer respostas prontas [Küchemann et al. 2025]. A eficácia da ferramenta depende, portanto, de sua capacidade de promover uma aprendizagem assistida e deve ser avaliada com base nos fatores que equilibram a usabilidade da interface, enquanto desenvolve competências para avaliar uma informação por conta própria [Joseph and Thomas 2024, Alabbas and Alomar 2025].

Diante desta lacuna, o presente trabalho apresenta o S.N.A.K.E. (*System for News Authentication and Knowledge Evaluation*), uma plataforma que propõe um modelo de interação de cunho mais pedagógico para o usuário, cujo objetivo é guiá-lo através do processo de análise, explicando os critérios de autenticidade e fomentando o letramento digital, transformando o ato da verificação em uma oportunidade de construção do pensamento crítico.

Ao combater a desinformação, a ferramenta auxilia no questionamento de notícias falsas, impulsionando o pensamento crítico sobre conteúdos enganosos que semeiam desconfiança nas instituições. Além disso, o projeto atua na construção de cidadãos conscientes e críticos no âmbito digital, formando uma população resiliente contra a exploração de vulnerabilidades cognitivas e a propagação acelerada de notícias falsas.

2. Fundamentação Teórica e Trabalhos Relacionados

A partir da contextualização anterior, é importante definir as questões referentes ao desenvolvimento de soluções computacionais para a desinformação, exigindo uma compreensão multidisciplinar [Praxedes 2024]. Neste sentido, a presente seção apresenta os conceitos teóricos que fundamentam a arquitetura e o propósito do S.N.A.K.E.

2.1. A Ameaça da Desinformação

O problema da desinformação vai além do ambiente virtual, manifestando-se como uma ameaça crítica à estabilidade democrática, à saúde pública e à coesão social [Oliveira and Gomes 2019, Praxedes 2024]. A circulação de notícias falsas demonstrou impactos graves na sociedade, como fomentar o declínio na cobertura vacinal contra o sarampo e a COVID-19 [Organização Pan-Americana da Saúde 2021]. Nesse contexto, vulnerabilidades cognitivas e o viés de confirmação dos usuários são situações exploradas pela disseminação da desinformação, que se propaga em velocidades até seis vezes maiores do que as informações corretas [Oliveira and Gomes 2019].

O combate a esse fenômeno é dificultado quando a propagação ocorre em plataformas de mensagens criptografadas de ponta a ponta, como o WhatsApp e o Telegram. Diferente das redes sociais abertas (e.g., X ou Instagram), nas quais a checagem de fatos por terceiros pode ser aplicada de maneira centralizada, nessas plataformas de mensagens a informação flui mediante conexões interpessoais (família, amigos, colegas de trabalho), o que gera, portanto, uma falsa aparência de verossimilhança ao conteúdo recebido [Frischlich et al. 2024].

Diante da inviabilidade de uma cautela externa nos ambientes privados, torna-se importante o desenvolvimento de soluções que capacitem o usuário a identificar manipulação e desinformação por si próprio. Assim, a transição para métodos que utilizem a IA como suporte ao letramento digital ativo se mostra como o caminho necessário para fortalecer a resiliência cognitiva do usuário, conforme discutido a seguir.

2.2. Abordagem por LLMs e Letramento Digital Ativo

O surgimento dos LLMs possibilitou a análise minuciosa não apenas da factualidade das informações, mas também de padrões linguísticos típicos da desinformação, como a polarização e o tom exagerado [Kuntur et al. 2025]. No cenário brasileiro, pesquisas recentes comprovam que esses modelos, incluindo versões de código aberto, atingem o estado da arte na detecção de notícias falsas em português [Gôlo et al. 2024]. Essa abordagem permite superar os tradicionais obstáculos relacionados à anotação manual de dados em larga escala [Gôlo et al. 2024]. Além disso, uma vantagem técnica decisiva dos LLMs reside na sua capacidade de explicabilidade. Dessa forma, a IA pode justificar os critérios linguísticos e contextuais que fundamentam a classificação de um conteúdo como enganoso.

Entretanto, o uso convencional de modelos de linguagem (isto é, quando a IA gera respostas com base em um único *prompt*) apresenta limitações para o propósito educativo. Para superar esse obstáculo, o S.N.A.K.E. propõe a utilização de *Agentic Workflows* [Li and Zhang 2024]. Nessa abordagem, o modelo deixa de ser uma ferramenta estática para emular as etapas de investigação de um *fact-checker* humano. O processo envolve a fragmentação da alegação principal, busca ativa por inconsistências lógicas e análise

da intenção do autor. A partir dessas etapas, o sistema constrói uma conclusão de forma iterativa e fundamentada [Li and Zhang 2024]. Com isso, a IA deixa de atuar como um classificador passivo e assume o papel de um par colaborativo no aprendizado assistido, promovendo o letramento digital e capacitando o usuário a analisar a estrutura das notícias [Joseph and Thomas 2024, Al-Zahrani 2024].

2.3. Trabalhos Relacionados e Inovação

No cenário nacional, a automação do *fact-checking* tem gerado iniciativas relevantes mediadas por aplicativos de mensagem. Como já visto anteriormente, destacam-se o *chatbot* "Fátima", desenvolvido pela agência Aos Fatos [Aos Fatos 2023], e o sistema "Tá Certo Isso Aí?" [Lira et al. 2025], que exemplificam o uso de LLMs para processar dúvidas do público e devolver checagens cruzadas de forma rápida. Outras frentes, como o Projeto Comprova [Projeto Comprova 2026] e o serviço "Fato ou Fake" do G1 [G1 2026], também operam canais de verificação, reforçando a tendência de utilizar a tecnologia para escalar o combate à desinformação.

Apesar da utilidade em tempo real, muitas soluções operam predominantemente sob uma lógica em que o usuário terceiriza o julgamento crítico à máquina e recebe um veredito de veracidade binário (e.g., 'Falso' ou 'Verdadeiro'), um fenômeno que estudos recentes identificam como descarregamento cognitivo [Fisher et al. 2026]. Ou seja, sem compreender os critérios da análise [Frischlich et al. 2024]. Consequentemente, a literatura indica que a dependência exclusiva de abordagens reativas é insuficiente para sanar o problema estrutural do pensamento crítico da população [Fisher et al. 2026], falhando quando o objetivo é construir uma resiliência de longo prazo contra a desinformação [Küchemann et al. 2025].

É nesta lacuna que o S.N.A.K.E. se posiciona, propondo uma transição do modelo de análise para uma arquitetura baseada na justificativa e no letramento ativo, de modo que vai além da funcionalidade de uma ferramenta puramente verificadora. Ao decompor os detalhes linguísticos e as táticas de manipulação presentes na notícia, o sistema atua como um tutor digital de letramento. O principal propósito é o fomento da participação do usuário na construção do raciocínio, desenvolvendo uma "imunidade cognitiva"[Roozenbeek et al. 2022, Bezzaoui et al. 2025].

3. Arquitetura e Funcionamento do Sistema S.N.A.K.E.

O S.N.A.K.E. prioriza a escalabilidade, utilizando uma comunicação assíncrona para neutralizar a latência referente às inferências de modelos de LLMs e às operações de *Retrieval-Augmented Generation* (RAG). A Figura 1 ilustra o fluxograma operacional do sistema, apresentando o pipeline de processamento que segmenta as responsabilidades entre segurança, análise léxica, extração semântica e síntese de resposta.

O ciclo de processamento inicia-se com o recebimento de *inputs* heterogêneos (texto bruto ou URLs). Imediatamente, os dados são submetidos a um *firewall* semântico que identifica ataques de *Prompt Injection* e filtra conteúdos que violam diretrizes éticas ou de privacidade. Uma camada do S.N.A.K.E. é dedicada a combater *spoofing* e *typosquatting*, empregando algoritmos de similaridade léxica (como a distância de Levenshtein) e mapeamento de caracteres Unicode para detectar ataques de homógrafos (substituição de caracteres latinos por variantes cirílicas ou gregas). A validação é

Figura 1. Fluxograma operacional do sistema S.N.A.K.E.

reforçada por um motor de *scoring* que cruza o domínio com *whitelists* de veículos jornalísticos e analisa *Top-Level Domains* (TLDs) de alta periculosidade, garantindo que o usuário seja notificado em caso de algum elemento suspeito no portal de notícias.

Em seguida, ocorrem o pré-processamento, a sanitização das entradas e a formulação de *queries* (para contexto externo), executados pelo modelo de baixa latência Sabiazinho-4[1]. O fluxo então avança para a etapa de Extração de Tópicos e Definição de Foco, na qual operações mais exigentes, como o reconhecimento de entidades nomeadas e a contextualização semântica, formam o núcleo de inteligência operado pelo SABIA-4.

Com o escopo informacional delimitado, o sistema inicia a fase de Investigação e Verificação. Esta etapa integra rotinas de *web scraping* em portais de notícias à arquitetura de RAG, utilizando as *queries* previamente construídas. Neste contexto, o modelo BERTimbau[2] é dedicado exclusivamente à geração de *embeddings*, garantindo precisão semântica na recuperação de documentos externos em português para o cruzamento de fatos. Após a coleta, os dados recuperados são submetidos à Avaliação Qualitativa, que atesta a consistência lógica entre as fontes externas e a afirmação original do usuário.

Consequentemente, a Geração de Resposta compõe uma devolutiva estruturada, integrando alertas de segurança gerados nas camadas iniciais e recomendações baseadas em evidências. Essa etapa é conduzida por diferentes modelos de *prompts* pedagógicos[3], responsáveis por orientar o LLM na explicação do raciocínio, na identificação de inconsistências informacionais e na sugestão de caminhos para a verificação da notícia.

O S.N.A.K.E. emprega técnicas de *prompt engineering* clássicas como *CO-STAR*, *Few-shot*, *Chain-of-Thought*, etc. De maneira complementar, foram desenvolvidos *prompts* agênticos que acionam ferramentas internas de análise linguística e factual, permitindo uma investigação sistemática e fundamentada [Li and Zhang 2024]. Por fim, o ciclo encerra-se na camada de Interação e Aprendizado, na qual o usuário pode continuar a discussão a partir das respostas fornecidas pelo sistema.

4. Resultados e Discussão

A ferramenta proposta é disponibilizada por meio de uma interface fluida, tanto na plataforma Web quanto via *bot* no Telegram. O diferencial operacional do S.N.A.K.E. é

[1]https://www.maritaca.ai/
[2]https://huggingface.co/neuralmind/bert-base-portuguese-cased
[3]Disponível em: https://snake.inteligentehub.com.br/

entregar uma análise fundamentada e contextualizada, distanciando-se das abordagens que se apoiam em validações puramente binárias. Na Figura 2, observa-se a recepção do conteúdo (*link* de notícia ou texto), no qual o sistema inicia o processamento da requisição do usuário. O fluxo resulta na geração de uma resposta estruturada que apresenta a validação do conteúdo, acompanhada de recomendações de tópicos para análise, a fim de incentivar o usuário a aprofundar sua investigação.

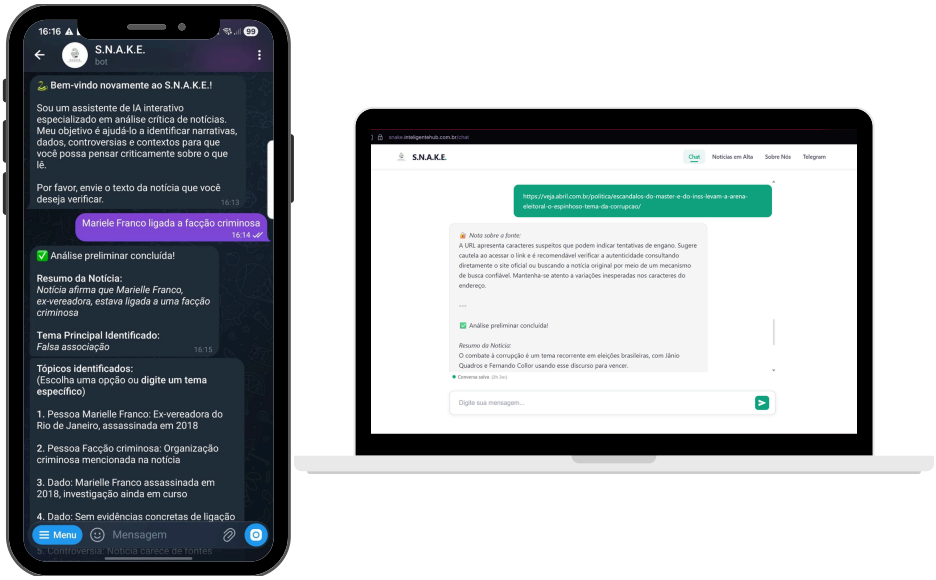

Figura 2. Interface de interação do usuário via Telegram e Web: exemplos de submissão de notícia e início do fluxo de processamento.

A Figura 3 ilustra o *output* educativo final, que explora alertas, recomendações e a justificativa do processo de análise dos pontos de discussão. Ao contrário de ferramentas que operam como ”caixas-pretas”, o S.N.A.K.E. atua como um tutor digital que explica os critérios de análise [Bezzaoui et al. 2025] e promove a capacitação do usuário na identificação de padrões de desinformação. Por meio desse diálogo interativo, o sistema demonstra as nuances e possíveis inconsistências lógicas presentes no tópico analisado, transformando o ato da verificação em uma oportunidade de construção de pensamento crítico.

4.1. Prova de Conceito

Para validar a proposta, o S.N.A.K.E. foi submetido a um conjunto de testes utilizando 40 notícias[4], balanceado entre conteúdos legítimos e enganosos. Como os sistemas avaliados não fornecem um veredito binário explícito, o desempenho foi calculado com base na análise das respostas geradas, utilizando indicadores linguísticos e semânticos. Como forma de comparação exploratória, o S.N.A.K.E. foi analisado em relação a duas plataformas de *fact-checking* e letramento midiático: o robô Fátima [Aos Fatos 2023] e o sistema Tá Certo Isso Aí [Lira et al. 2025], conforme apresentado na Tabela 1.

Para cada uma das notícias, extraiu-se da fonte original o trecho que sintetiza o fato verificado, preservando a literalidade do material consultado. As fontes contempladas

[4]Notícias e respostas dos testes: https://encurtador.com.br/Sosj - O número pequeno de notícias é justificado pelo limite de uso de algumas plataformas

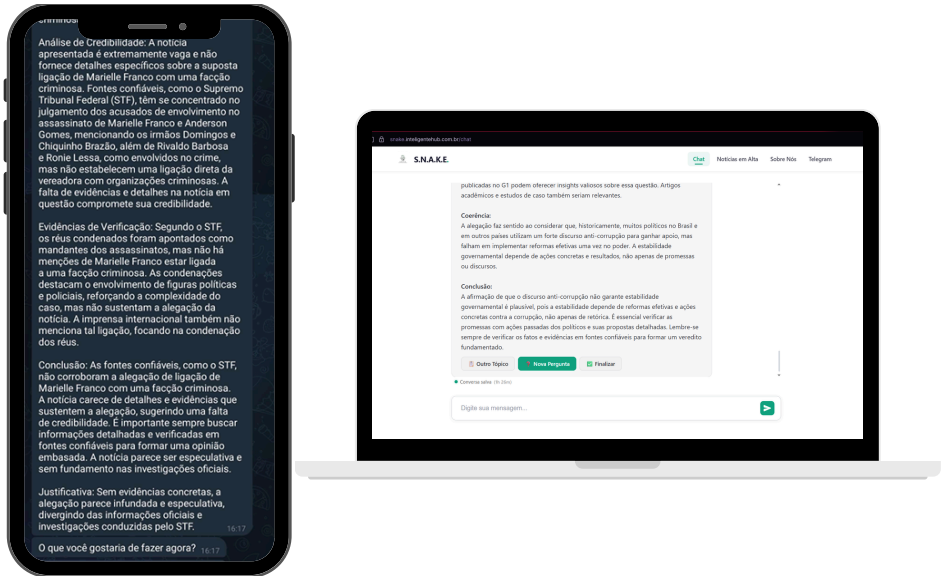

Figura 3. Devolutiva pedagógica do sistema: análise explicável destacando critérios linguísticos, evidências factuais e sugestões de letramento.

distribuem-se em quatro categorias: (i) agências brasileiras consolidadas de *fact-checking* (Aos Fatos), signatária verificada do código de princípios da *International Fact-Checking Network* (IFCN) desde 2017, sob auditoria anual independente; (ii) serviços editoriais de checagem de grandes grupos jornalísticos [G1 2026]; (iii) reportagens da imprensa nacional de referência (G1, Folha de S. Paulo, Estadão, etc); e (iv) fontes institucionais oficiais: pronunciamentos do STF, da Polícia Federal, de ministérios e de assessorias dos próprios atores mencionados na alegação.

Para avaliação, cinco métricas foram utilizadas: *Faithfulness* (Fidelidade à Evidência), estimada a partir da similaridade de cosseno entre a resposta produzida pelo sistema e as evidências factuais correspondentes, utilizando *embeddings* gerados pelo modelo BERTimbau. As métricas de *Evidence Attribution* e *Evidence Coverage* foram estimadas por meio da detecção automática de padrões lexicais associados a citações de fontes, instituições, dados e expressões de fundamentação, utilizando um dicionário expandido de termos para verificação de fatos. *Reasoning Depth*, que estima a profundidade argumentativa das respostas a partir da presença de estruturas lógico-discursivas associadas ao raciocínio passo a passo (*Chain-of-Thought*). Por fim, a métrica de *Misinformation Detection* é estimada pela presença de padrões linguísticos associados à problematização crítica da informação, como a identificação de inconsistências narrativas, linguagem alarmista ou ausência de fontes verificáveis.

Os resultados corroboram o perfil pedagógico e analítico do S.N.A.K.E., que obteve o desempenho mais expressivo em Profundidade de Raciocínio (3,464), um valor cerca de quatro vezes superior ao da segunda colocada. Essa pontuação reflete sua capacidade de encadear ideias de forma lógica, guiando o usuário pelos processos de compreensão e verificação. Também liderou as métricas de *Evidence Attribution* (2,834) e *Evidence Coverage* (1,681), apresentando aptidão para citar fontes, instituições e expressões de fundamentação de maneira incorporada à argumentação, em vez de se limitar à listagem mecânica de referências. Embora o robô Fátima tenha apresentado os maio-

Tabela 1. Comparativo de desempenho do S.N.A.K.E. frente a plataformas de *fact-checking* (métricas normalizadas por contagem de palavras).

Métrica	S.N.A.K.E.	Fátima	Tá Certo Isso Aí
Reasoning Depth	**3,464**	0,476	0,766
Evidence Coverage	**1,681**	1,138	0,661
Evidence Attribution	**2,834**	1,904	2,430
Misinformation Detection	0,597	**1,825**	0,478
Faithfulness (0–1)	0,803	**0,806**	0,783

res índices em *Misinformation Detection* (1,825) e *Faithfulness* (0,806), seu desempenho reduzido em *Reasoning Depth* (0,476) sugere respostas mais sintéticas, com menor encadeamento explicativo. Por fim, o S.N.A.K.E. manteve níveis elevados de *Faithfulness* (0,803), com diferença mínima em relação ao melhor resultado, indicando alinhamento semântico consistente com as evidências reais.

4.2. Avaliação de Usuários e Impacto Pedagógico

A avaliação preliminar do impacto pedagógico e da aceitação do sistema foi conduzida por meio de um questionário estruturado composto por 22 questões[5], aplicado a um grupo piloto de usuários ($N = 8$) com diferentes perfis em termos de idade, área de atuação e familiaridade tecnológica. O instrumento foi desenvolvido para capturar a percepção dos participantes sobre a experiência de uso do sistema. As questões adotam uma escala *Likert* de cinco pontos, variando de 1 (“Discordo totalmente”) a 5 (“Concordo totalmente”).

Um dos indicadores mais expressivos foi a disposição dos participantes em recomendar a ferramenta, com uma média de 4,63. Esse dado sugere que o sistema é percebido como uma camada de proteção socialmente necessária e confiável. No âmbito qualitativo, os usuários destacaram que o diferencial do sistema não reside na classificação em real ou *fake*, mas na capacidade de fornecer caminhos para a investigação autônoma. Relatos indicam que o fornecimento de fontes alternativas e sugestões de leitura sobre o mesmo tópico foi fundamental para a construção desse entendimento. Isso permite que o indivíduo decida de forma fundamentada sobre a confiabilidade de uma notícia.

Referente à usabilidade, com uma pontuação média de 4,63, os participantes concordaram que as respostas dos assistentes foram relevantes para o contexto da análise, enquanto 100% avaliaram a clareza das explicações entre ”claro” e ”muito claro” (4 e 5). Este dado é importante, pois a transparência no processo de análise é o que permite superar a desconfiança comum em sistemas de IA percebidos como classificadores que não fornecem justificativas.

Quanto ao impacto no letramento digital, os participantes afirmaram que a ferramenta ensinou algo novo sobre a identificação de notícias falsas, obtendo uma média de 3,38, enquanto o desenvolvimento de um olhar mais atento, através das dicas pedagógicas, alcançou 3,50. Embora um pouco inferiores aos índices de usabilidade, esses valores indicam que a interação breve com o sistema já inicia o processo de letramento [Roozenbeek et al. 2022].

[5]https://forms.gle/GyYXJqhdNUnmmMGP6

5. Limitações

As restrições operacionais das APIs/concorrentes delimitaram a prova de conceito a 40 notícias e a avaliação a um grupo piloto de 8 usuários, o que restringe a generalização estatística e caracteriza o estágio atual como uma validação inicial de viabilidade pedagógica. Além disso, o sistema pode herdar vulnerabilidades associadas aos modelos utilizados, como vieses em seus pesos, bem como limitações decorrentes da complexidade técnica da auditoria automatizada de fontes externas em tempo real. Finalmente, a ausência de testes formais de estresse sob alta concorrência ainda impede avaliar com precisão a latência, a estabilidade e o custo operacional em diferentes cenários.

6. Considerações Finais

O S.N.A.K.E. propõe a transição do paradigma de verificação binária para uma arquitetura agêntica de IA explicável voltada ao letramento digital sobre a desinformação. Nesse sentido, os achados indicam que o sistema cumpre, ainda que de forma inicial, seu propósito social ao fortalecer a autonomia crítica do usuário e capacitá-lo a navegar com maior segurança no ecossistema informacional. Diante das limitações identificadas nesta etapa exploratória, estabelecem-se as seguintes diretrizes para trabalhos futuros: (i) expansão da avaliação experimental para um corpus ampliado ($N \geq 100$); (ii) condução de estudos para mensurar a evolução e a retenção do olhar crítico dos usuários ao longo do tempo; (iii) otimização da infraestrutura; (iv) integração de modelos multimodais para detecção de desinformação em áudio e imagem.

Disponibilidade e Agradecimentos

O sistema e a sua documentação técnica, incluindo a biblioteca aberta de prompts pedagógicos para auditoria e replicação acadêmica, estão publicamente disponíveis em https://snake.inteligentehub.com.br/. Este trabalho foi financiado pelo *International Development Research Centre* (Canadá, Acordo 109981) e pelo *Foreign*, *Commonwealth and Development Office*.

Declaração do uso de IA Generativa

Os autores declaram que ferramentas de Inteligência Artificial Generativa (LLMs) foram utilizadas exclusivamente para melhorar a clareza e a qualidade textual do manuscrito.

Referências

Al-Zahrani, A. M. (2024). Exploring the impact of artificial intelligence on higher education: The dynamics of ethical, social, and educational implications. *Humanities and Social Sciences Communications*, 11(1):912.

Alabbas, A. and Alomar, K. (2025). A weighted composite metric for evaluating user experience in educational chatbots: Balancing usability, engagement, and effectiveness. *Future Internet*, 17(2).

Aos Fatos (2023). Fale com a robô checadora fátima para verificar informações e interagir com o conteúdo do aos fatos. `https://www.aosfatos.org/fatima/`. Acesso em: 21 jan. 2026.

Bezzaoui, I., Stein, C., Weinhardt, C., and Fegert, J. (2025). Explainable ai for online disinformation detection: Insights from a design science research project. *Electronic Markets*, 35(1):1–28.

Fisher, S. et al. (2026). Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *MDPI - Humanities*, 15(1):52–68.

Frischlich, L., Klapproth, J., Frank, S., Heckmann, M., Kunze, S. E., and Murgas, T. (2024). Fighting fakes on whatsapp—audience perspectives on fact bots as countermeasures. *The International Journal of Press/Politics*.

G1 (2026). Fato ou fake: O serviço de checagem do g1. `https://g1.globo.com/fato-ou-fake/`. Acesso em: 08 mar. 2026.

Gôlo, M. P. S., Mori, A. L. V., Oliveira, W. G., Barbosa, J. R., Graciano-Neto, V. V., de Lima, E. A., and Marcacini, R. M. (2024). On the use of large language models to detect brazilian politics fake news. In *Anais do Simpósio Brasileiro de Tecnologia da Informação e Linguagem Humana (STIL)*. SBC.

Joseph, A. and Thomas, A. (2024). Impact of digital literacy, use of ai tools and peer collaboration on ai assisted learning: Perceptions of the university students. *Digital Education Review*, (45). ISSN: 2013-9144.

Kuntur, S., Wróblewska, A., Paprzycki, M., and Ganzha, M. (2025). Under the influence: A survey of large language models in fake news detection. *IEEE Transactions on Artificial Intelligence*, 6(2):458–475.

Küchemann, S., Avila, K. E., Dinc, Y., Hortmann, C., Revenga, N., Ruf, V., Stausberg, N., Steinert, S., Fischer, F., Fischer, M., et al. (2025). On opportunities and challenges of large multimodal foundation models in education. *npj Science of Learning*, 10(1):11.

Li, X. and Zhang, Y. (2024). Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.

Lira, C. P., Costa, L. F. D., and Silva, P. H. F. (2025). Tá certo isso aí? `https://tacertoissoai.com.br`. Acesso em: 21 jan. 2026.

Oliveira, A. S. and Gomes, P. O. (2019). Os limites da liberdade de expressão: fake news como ameaça a democracia. *Revista de Direitos e Garantias Fundamentais*, 20(2):93–118.

Organização Pan-Americana da Saúde (2021). Pandemia de covid-19 leva a grande retrocesso na vacinação infantil, mostram novos dados da oms e unicef. Acesso em: 25 maio 2026.

Praxedes, W. (2024). Sociologia das redes sociais. *Revista Espaço Acadêmico*, (243). jan./fev./mar. 2024.

Projeto Comprova (2026). Projeto comprova: Jornalismo colaborativo contra a desinformação. `https://projetocomprova.com.br/`. Acesso em: 08 mar. 2026.

Roozenbeek, J., van der Linden, S., et al. (2022). Psychological inoculation of millions of social media users at scale. *Science Advances*, 8(34).