Relatório final de atividades
Projeto: SAVITS: Sistema de Avaliação e Validação de Impactos em Tecnologias Sociais

Coordenador do Projeto:	**Washington Luís Ribeiro de Carvalho Segundo**		
Nome do(a) pesquisador(a):	**Rafael Rocha**		
Vigência da Bolsa de Pesquisa:	**08/06/2026**	**Até:**	**07/09/2026**
Encerramento em:	**07/09/2026**		
Ref. Fundação:	**32725**	**Ref. Ibict:**	**008/2026-81**

Descrição das Atividades da Bolsa de Pesquisa:

Implementar modelos de processamento de linguagem natural (NLP) para análise semântica, classificação de textos e extração de entidades nos dados qualitativos do Caso Piloto do SAVITS; Investigar e aplicar técnicas de RAG, embeddings semânticos e busca vetorial para suporte às funcionalidades de recuperação e geração de informações dos módulos LLM do SAVITS; Analisar requisitos técnicos e metodológicos de interoperabilidade e APIs abertas, sistematizando diretrizes para integração experimental dos componentes de NLP e avaliando métricas de desempenho em ambiente controlado; Elaborar relatório técnico final de bolsa, consolidando resultados, atividades desenvolvidas e contribuições metodológicas da pesquisa.

1 INTRODUÇÃO

O Sistema de Avaliação e Validação de Impactos em Tecnologias Sociais (SAVITS) demanda mecanismos computacionais capazes de processar, organizar e recuperar informações provenientes de dados qualitativos heterogêneos (IBICT, 2026). Nesse contexto, técnicas de *Natural Language Processing* (NLP) apresentam potencial para apoiar a análise semântica (DEVLIN et al., 2019); a classificação automática de textos e a extração de entidades, facilitando a identificação de padrões, temas e relações relevantes nos dados do Caso Piloto.

Além dessas técnicas, o uso de *embeddings* semânticos e mecanismos de busca vetorial permite representar e recuperar conteúdos com base em sua proximidade conceitual, reduzindo as limitações de abordagens exclusivamente orientadas por palavras-chave. Quando integrados a arquiteturas de Retrieval-Augmented Generation (RAG), esses recursos podem fornecer aos *Large Language Model* (LLM) contextos recuperados de bases previamente selecionadas (GAO et al., 2023), contribuindo para a geração de respostas mais fundamentadas, rastreáveis e aderentes às fontes disponíveis.

Diante desse cenário, esta pesquisa tem como propósito investigar a aplicação de modelos de NLP aos dados qualitativos do Caso Piloto do SAVITS, contemplando tarefas de análise semântica, classificação textual e extração de entidades. Busca-se, ainda, analisar e experimentar técnicas de RAG, embeddings semânticos e busca vetorial para apoiar as funcionalidades de recuperação e geração de informações dos módulos baseados em LLM.

Espera-se que os resultados contribuam para o desenvolvimento de mecanismos mais transparentes, rastreáveis e semanticamente fundamentados de análise dos registros produzidos no âmbito das Tecnologias Sociais. Ao combinar métodos de PNL, recuperação semântica e princípios de ciência cidadã, a pesquisa pretende apoiar a transformação de dados qualitativos em evidências úteis à avaliação de impactos,

preservando a centralidade da interpretação humana, o respeito às comunidades participantes e o caráter não vinculante das inferências computacionais. Ao final, as atividades, os resultados experimentais e as contribuições metodológicas serão consolidados em relatório técnico, de modo a subsidiar o aperfeiçoamento dos módulos de inteligência artificial e da infraestrutura informacional do SAVITS.

2 MÉTODO

A pesquisa desenvolve-se por meio de uma abordagem aplicada, exploratória e experimental (GIL, 2002). Utilizou o método da revisão integrativa da literatura, para levantar evidências sobre a temática estudada. Além de análise dos dados qualitativos sintéticos[1] do Caso Piloto do SAVITS e avaliação comparativa de modelos, técnicas e arquiteturas computacionais. A metodologia será organizada em etapas sucessivas e complementares, de modo a subsidiar tanto a seleção das tecnologias quanto a validação de sua adequação às necessidades do sistema.

2.1 Revisão integrativa da literatura

O percurso metodológico foi estruturado com o objetivo de garantir maior sistematicidade, transparência e reprodutibilidade ao processo de identificação, seleção e análise dos estudos relacionados ao uso da inteligência artificial no contexto das tecnologias sociais. Para isso, foi definido um protocolo composto por etapas sucessivas, abrangendo desde a formulação da questão de pesquisa até a análise dos trabalhos selecionados.

O protocolo metodológico foi organizado nas seguintes etapas:

Etapa 1 - Formulação da questão de pesquisa: definiu-se a seguinte questão: como a inteligência artificial está sendo utilizada no contexto das tecnologias sociais?

1 Como os dados do SAVITS ainda não haviam sido submetidos aos procedimentos de triagem e obfuscação necessários à proteção de informações potencialmente sensíveis, optou-se pela utilização de dados qualitativos sintéticos para a realização e o prosseguimento das análises experimentais.

Etapa 2 - Seleção das bases de dados: a Web of Science (WoS) foi selecionada por reunir publicações de milhares de editoras e adotar elevados critérios de indexação. A Base de Dados em Ciência da Informação (BRAPCI) também foi consultada, porém não retornou registros relacionados à estratégia de busca estabelecida. Por essa razão, a revisão considerou apenas os resultados recuperados na WoS. O acesso à base foi realizado por meio do Portal de Periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), utilizando o acesso remoto via Comunidade Acadêmica Federada (CAFe).

Etapa 3 - Elaboração da estratégia de busca: foram selecionados os termos “Social Technologies” e “IA”. A *string* de busca utilizada foi: (“Social Technologies” AND “IA”).

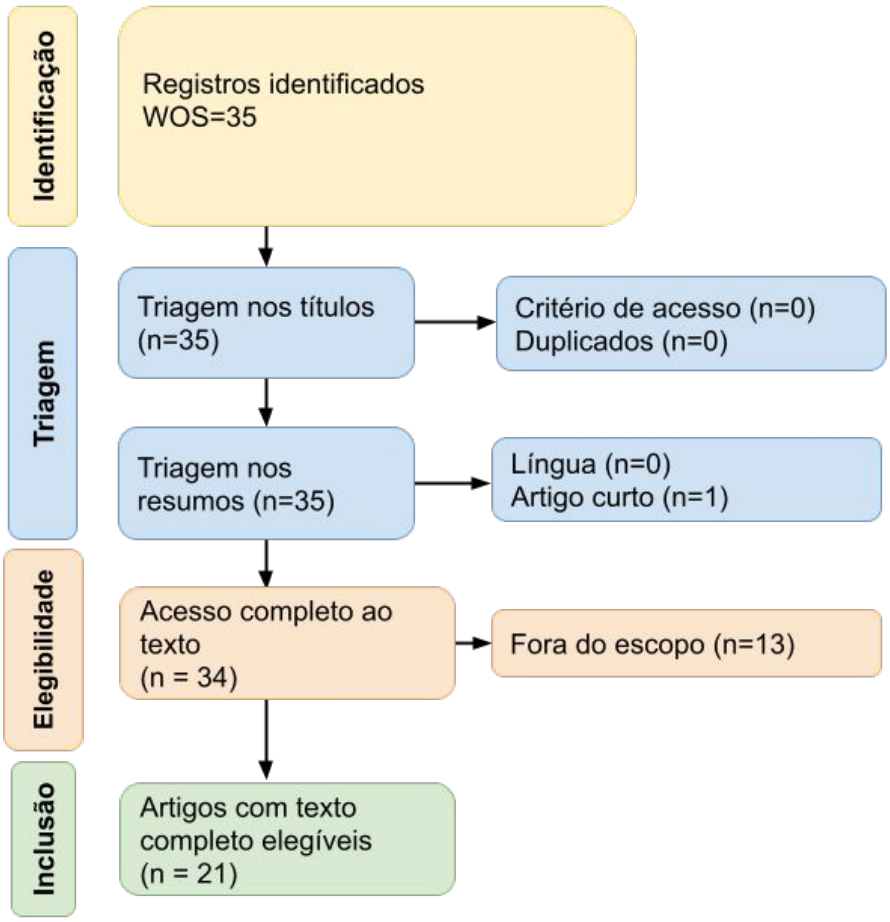
Etapa 4 - Definição dos critérios de inclusão e exclusão: foram incluídas publicações em inglês, português ou espanhol, publicadas entre 2025 e 2026, com acesso online ao texto completo e que apresentassem aplicação da temática investigada, com corpus claramente definido e análise objetiva dos textos. Foram excluídos estudos duplicados, artigos com menos de cinco páginas e trabalhos redundantes. Nos casos de redundância, foi considerada apenas a versão mais completa.

Etapa 5 - Análise dos estudos selecionados: os trabalhos incluídos foram examinados de acordo com sua aderência à questão de pesquisa e aos critérios previamente estabelecidos.

A aplicação dessas etapas possibilitou organizar o processo de revisão de maneira sequencial, desde a delimitação do problema investigado até a constituição do conjunto final de estudos. A seleção das publicações foi conduzida com base nos critérios definidos no protocolo, buscando manter a correspondência entre os trabalhos recuperados, a temática das tecnologias sociais e a aplicação da inteligência artificial.

Por fim, a Figura 1 apresenta graficamente o processo adotado na revisão integrativa da literatura, sintetizando as etapas de busca, seleção, aplicação dos critérios de elegibilidade e análise dos estudos incluídos.

Figura 1 - Fluxograma PRISMA do processo de inclusão e extração.



Fonte: Dados da pesquisa (2026)

Os resultados obtidos a partir da aplicação dos procedimentos metodológicos descritos serão apresentados, analisados e discutidos na Seção 3, na qual serão examinados à luz dos objetivos estabelecidos para a pesquisa.

2.2 Benchmark de modelos de linguagem

Com o objetivo de identificar modelos de linguagem com maior aderência potencial às aplicações em tecnologias sociais, foi realizado um *benchmark* comparativo de LLMs. A análise considerou informações disponibilizadas em rankings, conjuntos de

resultados de avaliação e fichas técnicas de modelos presentes nas plataformas Hugging Face[2] e LMArena[3]. A Hugging Face reúne avaliações, leaderboards comunitários e model cards que documentam características, resultados, capacidades e limitações dos modelos. A LMArena, por sua vez, compara modelos a partir de avaliações humanas realizadas por meio de confrontos pareados e utiliza métodos estatísticos para consolidar as preferências observadas em seus rankings.

A utilização conjunta dessas fontes permitiu combinar duas perspectivas complementares. Os dados provenientes da Hugging Face foram empregados principalmente na identificação de resultados obtidos em tarefas padronizadas, características técnicas, condições de licenciamento e informações declaradas pelos desenvolvedores. Os dados da LMArena foram utilizados como indicadores da qualidade percebida em interações mais abertas, aproximando a análise de situações reais de uso. Essa combinação buscou evitar que a seleção dos modelos se apoiasse exclusivamente em um único *benchmark* ou em resultados produzidos sob condições experimentais específicas.

Inicialmente, foram identificados os modelos com resultados disponíveis e informações suficientes para sua comparação. Em seguida, os dados foram organizados em seis dimensões: capacidade de linguagem, seguimento de instruções, análise de dados, raciocínio, custo e licença de uso. Sempre que uma dimensão foi representada por mais de uma métrica, os resultados correspondentes foram agregados para a obtenção de uma pontuação única por categoria.

Para possibilitar a comparação entre métricas expressas em escalas distintas, os valores foram normalizados para uma escala comum. Nos indicadores em que valores mais elevados representavam melhor desempenho, a normalização preservou essa relação. Nos critérios de custo, a escala foi invertida, de modo que modelos com menor custo relativo recebessem pontuações superiores. Nos casos de ausência de

2 Disponível em: https://huggingface.co acesso em 30/07/2026.
3 Disponível em: https://arena.ai acesso em 30/07/2026.

informações, o dado foi registrado como não disponível, evitando a atribuição de desempenho não demonstrado.

A pontuação final de cada modelo foi calculada por meio de um índice ponderado, denominado Índice de Mensuração do Potencial de Aderência e Conformidade às Tecnologias Sociais (IMPACTS)[4]. Os pesos foram definidos de acordo com a relevância prática de cada dimensão para aplicações sociais e institucionais, conforme apresentado no Quadro 1.

Quadro 1 - Pesos do IMPACTS

Dimensão avaliada	Peso
Capacidade de linguagem	25%
Seguimento de instruções	25%
Análise de dados	20%
Raciocínio	15%
Custo	10%
Licença aberta	5%

Fonte: Autor (2026)

A capacidade de linguagem recebeu peso de 25%, considerando que as tecnologias sociais dependem de comunicação clara, acessível e contextualizada em atividades como atendimento, educação, orientação, produção de documentos e interação com diferentes públicos. Nessa dimensão, foram considerados indicadores relacionados à compreensão e à geração de linguagem, com atenção ao desempenho em português e em tarefas de comunicação mais abertas.

4 O Índice de Mensuração do Potencial de Aderência e Conformidade às Tecnologias Sociais (IMPACTS) foi formulado inicialmente com base em uma heurística de caráter exploratório, derivada da experiência acumulada do pesquisador e da literatura analisada. Reconhece-se, contudo, que sua estrutura e ponderação demandam validação empírica posterior, especialmente por meio de grupos focais, entrevistas com especialistas e outros procedimentos de avaliação participativa.

O seguimento de instruções também recebeu peso de 25%, pois parte significativa das aplicações institucionais ocorre dentro de protocolos, formulários, políticas, normas e fluxos de trabalho previamente definidos. Dessa forma, avaliou-se a capacidade dos modelos de cumprir comandos, respeitar restrições e produzir respostas em formatos determinados.

A análise de dados recebeu peso de 20%, em razão da necessidade de processar diagnósticos, relatórios, pesquisas, indicadores, documentos e relatos qualitativos. Essa dimensão contemplou a capacidade de interpretar informações, identificar relações, sintetizar conteúdos e extrair elementos relevantes de diferentes tipos de registro.

O raciocínio recebeu peso de 15%, por contribuir para a interpretação de situações, a comparação de alternativas e o apoio à tomada de decisão. A atribuição de um peso inferior ao das dimensões de linguagem e seguimento de instruções decorre do entendimento de que os modelos devem atuar como instrumentos de apoio, sem substituir a avaliação humana em decisões sensíveis ou socialmente relevantes.

O custo correspondeu a 10% da pontuação final. Esse critério foi incorporado porque a viabilidade financeira influencia diretamente a adoção de modelos por organizações públicas, comunitárias e sociais com recursos limitados. Entretanto, seu peso foi mantido abaixo dos critérios de desempenho, de modo que um menor custo não compensasse limitações significativas de qualidade ou utilidade.

A licença aberta recebeu peso de 5%. Essa dimensão valorizou modelos que possibilitam maior transparência, auditoria, adaptação local e autonomia tecnológica. A ponderação reconhece a abertura como um diferencial estratégico, sem excluir modelos proprietários que possam apresentar utilidade prática para determinadas aplicações.

A pontuação ponderada foi obtida pela soma dos resultados normalizados de cada dimensão multiplicados pelos respectivos pesos. Assim, o índice foi calculado pela seguinte expressão:

$$IMPACTS = 0{,}25L + 0{,}25I + 0{,}20D + 0{,}15R + 0{,}10C + 0{,}05A$$

Em que no IMPACTS; (L), a pontuação de linguagem; (I), o seguimento de instruções; (D), a análise de dados; (R), o raciocínio; (C), o custo; e (A), a abertura da licença.

Desse modo, 85% da pontuação correspondeu às capacidades operacionais dos modelos, enquanto os 15% restantes incorporaram fatores relacionados à viabilidade de adoção e à autonomia tecnológica. A classificação resultante não buscou determinar o melhor modelo de maneira absoluta, mas identificar aqueles que apresentavam maior compatibilidade potencial com soluções voltadas à educação, à saúde, à assistência social, à participação cidadã e aos serviços públicos.

A expressão “aderência potencial” foi adotada porque os resultados do *benchmark* indicam possibilidades de aplicação, e não garantias de desempenho em ambientes reais. A adequação efetiva de cada modelo permanece condicionada à validação com dados do domínio, à proteção das informações, às práticas de governança, à avaliação de riscos e à supervisão humana. Além disso, como rankings e resultados de avaliação podem ser atualizados ao longo do tempo, foram considerados os dados disponíveis no período de realização da coleta.

Por fim, os resultados serão apresentados na Seção 3.2 onde foram organizados em uma classificação comparativa, acompanhada das pontuações obtidas em cada dimensão. Essa apresentação oferece subsídios para a seleção de alternativas compatíveis com os requisitos tecnológicos, sociais e institucionais do SAVITS.

2.3 Benchmark com dados qualitativos sintéticos

Foi realizado um *benchmark* experimental de LLMs com o objetivo de avaliar seu desempenho em tarefas representativas de aplicações de inteligência artificial no contexto das tecnologias sociais. Para essa finalidade, utilizou-se um conjunto composto

por 40 casos qualitativos sintéticos[5], elaborados para reproduzir situações plausíveis de atuação social e institucional, sem utilizar dados pessoais reais.

O conjunto contemplou dez domínios baseados nos resultados obtidos na Seção 3.1: (i) saúde comunitária; (ii) educação e inclusão digital; (iii) agricultura familiar; (iv) gestão de resíduos; (v) água e saneamento; (vi) segurança alimentar; (vii) habitação e urbanismo social; (viii) trabalho e geração de renda; (ix) participação cidadã; e (x) prevenção de riscos e desastres. Para cada domínio, foram elaboradas quatro tarefas, totalizando dez casos para cada uma das seguintes categorias: linguagem e simplificação; seguimento de instruções; análise de dados; e raciocínio e recomendação.

A seleção dessas tarefas buscou representar capacidades necessárias à aplicação de LLMs em tecnologias sociais. A tarefa de linguagem e simplificação avaliou a capacidade de produzir conteúdos claros, acessíveis e adequados ao público indicado. O seguimento de instruções examinou o cumprimento de restrições de formato, conteúdo e extensão. A análise de dados verificou a interpretação de informações, a realização de cálculos e o reconhecimento dos limites dos dados apresentados. Por fim, a tarefa de raciocínio e recomendação avaliou a formulação de propostas coerentes, acompanhadas de justificativas, riscos e medidas de mitigação.

Cada caso foi constituído por um problema, dados de entrada, público destinatário, prompt de teste, resposta de referência e critérios de avaliação. As respostas produzidas pelos modelos foram comparadas com os requisitos definidos em cada caso, considerando aspectos como clareza, fidelidade ao conteúdo, cumprimento das instruções, correção dos cálculos, coerência da justificativa e ausência de informações não fundamentadas.

A construção do *benchmark* também incorporou princípios relacionados às tecnologias sociais, como participação comunitária, acessibilidade, baixo custo, não discriminação, proteção de dados e respeito à autonomia dos participantes. Esses

5 Os casos analisados encontram-se disponíveis em repositório público, permitindo sua consulta e verificação: https://gist.github.com/rafaro/c6c5a5345aa968442497c63d9620ec41.

elementos foram incluídos nos prompts e nos critérios de avaliação para verificar se os modelos eram capazes de preservar tais requisitos durante a produção das respostas.

A avaliação comparativa das respostas geradas pelos modelos[6] ChatGPT e Claude foi realizada por meio de comparações pareadas, utilizando o identificador de cada questão como chave de correspondência entre as respostas produzidas e o respectivo texto de referência. Para cada ID, foram calculadas três métricas complementares de similaridade; similaridade sequencial, cosseno lexical e índice de Jaccard, além de um índice médio obtido pela agregação dos resultados dessas métricas.

A seleção desses algoritmos buscou contemplar diferentes dimensões da proximidade textual. A similaridade sequencial (RATCLIFF e METZENER, 1988) foi utilizada para verificar a correspondência entre as cadeias de caracteres, considerando a ordem em que palavras e trechos aparecem. Essa abordagem é sensível à estrutura da resposta e permite identificar reproduções, reorganizações limitadas e sequências textuais próximas ao gabarito. Sua inclusão mostra-se relevante em situações nas quais a formulação e o encadeamento dos conteúdos são importantes para a precisão da resposta.

O cosseno lexical (JONES e MEWHORT, 2007) foi empregado para mensurar a proximidade entre os textos a partir de suas representações vetoriais. Nessa abordagem, as respostas são representadas segundo a frequência ou a importância dos termos, e a similaridade é calculada pelo ângulo entre os vetores correspondentes. A métrica privilegia a proximidade vocabular e temática, sendo menos afetada pela extensão dos textos e por alterações na ordem das palavras. Dessa forma, permite verificar se a resposta do modelo mobiliza vocabulário e conceitos semelhantes aos presentes no texto de referência.

6 Por limitações orçamentárias, os experimentos foram conduzidos com as versões dos modelos disponíveis nas respectivas interfaces de conversação. Para o ChatGPT, utilizou-se o plano gratuito, cuja documentação não assegura a identificação permanente do modelo subjacente. Para o Claude, empregou-se o modelo Claude Sonnet 5, configurado no nível de esforço Extra, correspondente à maior intensidade de processamento disponibilizada na interface durante a realização do experimento.

O índice de Jaccard (BOUCHARD et al., 2013), por sua vez, avaliou a proporção de elementos lexicais compartilhados entre a resposta gerada e o gabarito. O índice corresponde à razão entre a interseção e a união dos conjuntos de termos dos dois textos. Essa métrica oferece uma interpretação direta da sobreposição vocabular e reduz a influência de repetições excessivas, uma vez que considera predominantemente a presença ou ausência dos elementos analisados.

As três métricas foram adotadas de forma complementar porque nenhuma delas, isoladamente, representa integralmente a qualidade de uma resposta textual. Enquanto a similaridade sequencial enfatiza a organização e a ordem do conteúdo, o cosseno lexical avalia a proximidade global entre as representações dos textos, e o índice de Jaccard evidencia a quantidade relativa de termos compartilhados. O índice médio foi utilizado como medida sintética, reduzindo a dependência de um único critério e possibilitando uma comparação global entre os modelos. Neste contexto, os resultados serão apresentados na seção a seguir. Os resultados são discutidos em profundidade na Seção 3.3.

3 RESULTADOS

Esta seção apresenta os resultados obtidos nas diferentes etapas da pesquisa, articulando as evidências provenientes da revisão integrativa da literatura, da análise de aderência dos modelos de linguagem de grande escala às tecnologias sociais e da experimentação conduzida com dados do domínio. Em conjunto, os achados permitem compreender como a inteligência artificial vem sendo incorporada às tecnologias sociais, quais modelos apresentam características mais compatíveis com esse campo e em que medida suas respostas reproduzem os conteúdos esperados segundo os critérios estabelecidos no estudo.

A Seção 3.1 apresenta os resultados da revisão integrativa da literatura, destacando as principais aplicações, tendências e lacunas identificadas nos estudos selecionados. Na sequência, a Seção 3.2 examina as LLMs com maior potencial de

aderência às tecnologias sociais, considerando os critérios e indicadores definidos na pesquisa. Por fim, a Seção 3.3 reúne os resultados da experimentação, apresentando o desempenho comparativo dos modelos e discutindo as diferenças observadas entre as respostas produzidas.

3.1 Resultados da revisão integrativa da literatura

A aplicação dos critérios de elegibilidade resultou na seleção preliminar de 21 estudos, publicados entre 2025 e 2026. A análise dos títulos, resumos e palavras-chave permitiu agrupá-los em seis eixos: trabalho e bem-estar; educação e cultura; saúde e cuidado; democracia e setor público; governança e sustentabilidade; e impactos sociotécnicos da inteligência artificial.

Embora a maioria dos trabalhos não utilize explicitamente a expressão “tecnologia social”, os estudos selecionados abordam aplicações, formas de adoção e impactos da inteligência artificial relacionados a problemas coletivos, serviços públicos, bem-estar, participação social e transformação institucional. Assim, a IA foi analisada como componente de arranjos sociotécnicos com efeitos sobre indivíduos, organizações e comunidades.

No eixo “trabalho, organizações e bem-estar”, Sun e Gao (2025) investigaram os efeitos da automação inteligente sobre a satisfação profissional e o bem-estar subjetivo, demonstrando que seus impactos variam conforme o vínculo empregatício e as condições de trabalho. Del Percio (2025) analisou discursos sobre IA, recrutamento e futuro do trabalho, evidenciando que narrativas tecnofuturistas podem apresentar a automação como solução inevitável, ocultando interesses econômicos e relações de poder. Dinis (2026) discutiu a adoção de sistemas inteligentes em funções gerenciais, defendendo uma articulação entre capacidades humanas e computacionais. Cuesta-Valino et al. (2026), por sua vez, relacionaram transformação digital, liderança e inteligência artificial à construção de ambientes organizacionais orientados ao bem-

estar. Esses estudos indicam que os benefícios da IA no trabalho dependem de políticas de proteção, supervisão humana e modelos de gestão socialmente responsáveis.

No campo da “educação e da cultura”, Wang et al. (2026) analisaram fatores que influenciam a intenção de futuros professores de utilizar IA em atividades pedagógicas, destacando a importância do conhecimento tecnológico, pedagógico e das condições institucionais. Goto e Ramnarain (2026) demonstraram que elementos culturais, influência social e disposição individual interferem na adoção da IA no ensino superior. Wang e Zhang (2026) compararam visitas educativas a museus conduzidas por sistemas de IA e por guias humanos, identificando diferenças cognitivas e socioafetivas entre as duas formas de mediação. Em conjunto, os resultados sugerem que a efetividade educacional da IA depende da formação dos usuários, do contexto cultural e da integração entre recursos automatizados e interação humana.

Os estudos de “saúde, cuidado e autonomia” enfatizaram a centralidade do usuário e a necessidade de explicabilidade. Francese et al. (2025) apresentaram aplicações de IA e interação humano-computador voltadas à reabilitação, ao diagnóstico, à saúde mental e ao cuidado de pessoas com demência. O trabalho destacou a personalização e a autonomia dos pacientes como princípios essenciais. Semel (2026) analisou criticamente sistemas que utilizam sinais vocais para identificar sofrimento mental, mostrando que sua aparente objetividade depende de decisões humanas sobre seleção e interpretação de dados. Fadillah e Jandevi (2026) discutiram a empatia sintética e a transformação do cuidado pela mediação computacional. Esses estudos evidenciam o potencial da IA para ampliar o acesso e a personalização dos serviços, mas também alertam para o risco de reduzir experiências humanas complexas a categorias algorítmicas.

No eixo “democracia, informação e setor público”, Vacha (2025) discutiu o potencial da IA para ampliar serviços públicos, interação entre Estado e sociedade e participação política, ao mesmo tempo que destacou riscos relacionados à concentração de dados. Ruvalcaba-Gomez e Garcia-Benitez (2026) compararam políticas públicas de

IA na Espanha e no México, identificando diferenças nos níveis de institucionalização e governança. Herrero de la Puente, Sancho Belinchon e Sedeno Lopez (2025) analisaram conteúdos falsos divulgados em contexto eleitoral e propuseram uma ferramenta baseada em IA para apoiar a identificação de desinformação. Becker (2026) examinou os impactos da IA generativa sobre autoria, legitimidade e responsabilidade institucional. Os trabalhos indicam que a IA pode fortalecer serviços e processos democráticos, desde que acompanhada por transparência, responsabilização e supervisão humana.

As pesquisas sobre "governança, regulação e sustentabilidade" demonstraram que os efeitos da IA ultrapassam sua dimensão técnica. Trefler (2026) apresentou argumentos econômicos para sua regulação, enfatizando riscos de concentração de mercado e aumento das desigualdades. Braun e Zanotti (2026) abordaram a IA como tecnologia ambivalente, capaz de produzir benefícios e danos simultaneamente. Dutu-Buzura (2026) diferenciou o uso da IA em favor do meio ambiente da necessidade de tornar os próprios sistemas computacionais mais sustentáveis. Silva (2025) analisou estratégias latino-americanas de IA e mostrou contradições entre propostas de integração regional e estruturas persistentes de dependência tecnológica. Os resultados reforçam que a apropriação social da IA depende de governança pública, sustentabilidade e autonomia tecnológica.

No eixo dos "impactos sociotécnicos", Lu (2026) investigou desigualdades digitais e mostrou que diferenças socioeconômicas influenciam o conhecimento e a capacidade de apropriação da IA. Nielsen (2025) analisou experimentalmente a percepção de privacidade em situações de vigilância humana e automatizada, destacando implicações para políticas de proteção de dados. Zajko (2026) discutiu os riscos de atribuir autonomia excessiva a sistemas inteligentes e defendeu a preservação da responsabilidade humana. Esses estudos demonstram que as desigualdades relacionadas à IA envolvem não apenas acesso material, mas também capacidade de compreender, contestar e influenciar os sistemas.

De forma geral, os estudos selecionados evidenciam quatro formas principais de utilização social da inteligência artificial: apoio a serviços de saúde, educação, cultura e administração pública; mediação de decisões e processos informacionais; transformação de instituições e políticas; e análise crítica de riscos relacionados ao trabalho, à privacidade, à desigualdade e à sustentabilidade.

Os resultados indicam que a IA pode contribuir para a resolução de problemas sociais quando orientada à autonomia dos usuários, ao interesse público, à ampliação do acesso e ao bem-estar coletivo. Entretanto, sua aderência aos princípios das tecnologias sociais depende de fatores como participação dos grupos envolvidos, transparência, explicabilidade, proteção de direitos, sustentabilidade e supervisão humana.

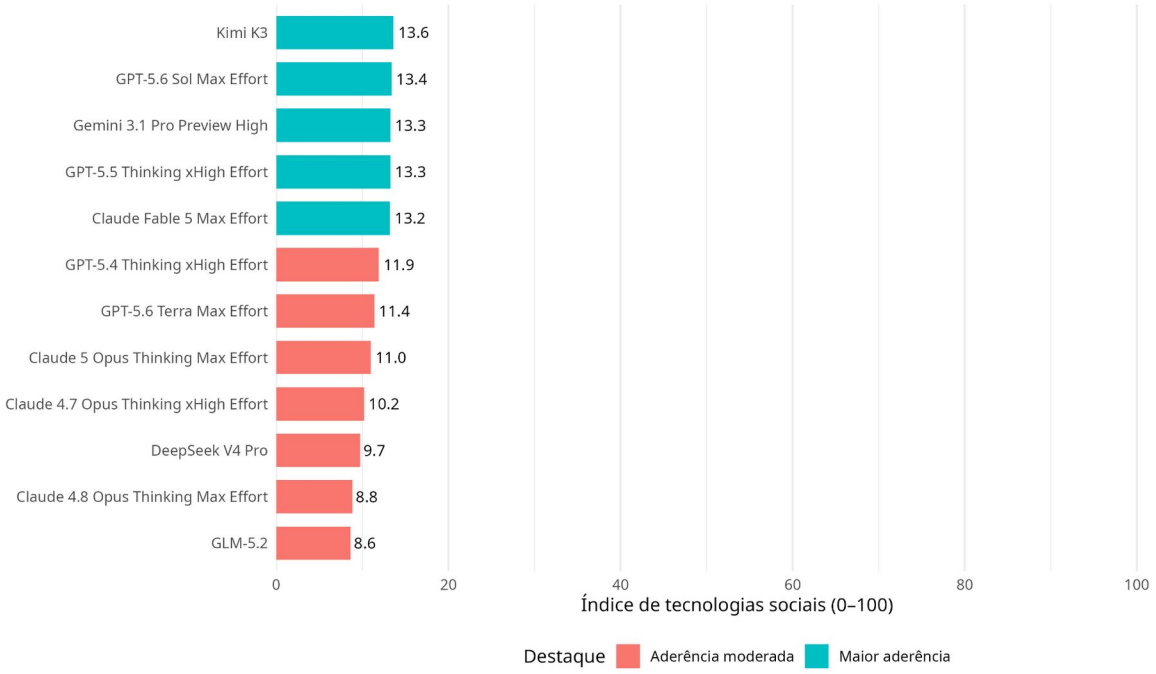
Apesar da proximidade temática, poucos estudos articulam explicitamente a inteligência artificial ao referencial das tecnologias sociais, especialmente aos princípios de construção participativa, apropriação comunitária, replicabilidade e transformação social. Essa ausência revela uma lacuna de pesquisa e sugere a necessidade de estudos que integrem de forma mais direta os recursos da IA às metodologias e aos valores das tecnologias sociais.

3.2 LLMs com maior aderência às tecnologias sociais

A aplicação do índice exploratório permitiu comparar a aderência potencial dos modelos de linguagem a aplicações relacionadas às tecnologias sociais. O indicador reuniu as dimensões de linguagem, seguimento de instruções, análise de dados e raciocínio, considerando também o custo[7] por tarefa bem-sucedida e o grau de abertura dos modelos. A Figura 2 apresenta a ordenação dos modelos conforme o índice agregado.

Figura 2 - Modelos com maior aderência potencial a tecnologias sociais

7 Esse indicador considera indiretamente o volume de tokens consumidos e a efetividade do modelo, distinguindo-se do preço nominal por mil tokens, que representa apenas o custo unitário de processamento.



Fonte: Dados da pesquisa (2026)

Já o Quadro 2 detalha os resultados obtidos em cada dimensão e a classificação atribuída.

Quadro 2 - Modelos relevantes para aplicações em tecnologias sociais

Modelo	Tipo	Índice social	Linguagem	Instruções	Dados	Raciocínio	Custo (US$)	Classificação
Kimi K3	Aberto	13,6	85,5	71,4	78,7	90,7	0,348	Maior aderência
GPT-5,6 Sol Max Effort	Proprietário	13,4	87,7	71,8	79,8	91,7	0,438	Maior aderência
Gemini 3,1 Pro Preview High	Proprietário	13,3	85,4	79,1	78,5	84	0,286	Maior aderência
GPT-5,5 Thinking xHigh Effort	Proprietário	13,3	87,4	70,7	81,6	89,7	0,435	Maior aderência
Claude Fable 5 Max	Proprietário	13,2	90,7	75,8	80,5	89,7	1,439	Maior

Effort								aderência
GPT-5,4 Thinking xHigh Effort	Proprietário	11,9	82,6	70,2	79,3	88,1	0,387	Aderência moderada
GPT-5,6 Terra Max Effort	Proprietário	11,4	82,9	64,6	79,3	90,6	0,352	Aderência moderada
Claude 5 Opus Thinking Max Effort	Proprietário	11	88,7	63,8	74,6	91,2	0,699	Aderência moderada
Claude 4,7 Opus Thinking xHigh Effort	Proprietário	10,2	77,9	66,7	78,3	87,2	0,528	Aderência moderada
DeepSeek V4 Pro	Aberto	9,7	78,1	62,4	74,5	82,7	0,050	Aderência moderada
Claude 4,8 Opus Thinking Max Effort	Proprietário	8,8	79,7	72	66	89,2	0,983	Aderência moderada

Fonte: Dados da pesquisa (2026)

Os maiores resultados foram alcançados pelo Kimi K3, com índice de 13,6, seguido pelo GPT-5.6 Sol Max Effort, com 13,4. O Gemini 3.1 Pro Preview High e o GPT-5.5 Thinking xHigh Effort obtiveram 13,3, enquanto o Claude Fable 5 Max Effort alcançou 13,2. A pequena diferença entre esses valores demonstra que os cinco modelos classificados com maior aderência apresentaram desempenho agregado relativamente próximo, embora tenham alcançado suas posições por meio de diferentes combinações de capacidades técnicas, custo e abertura.

O Kimi K3 ocupou a primeira posição e foi o único modelo aberto entre os cinco primeiros colocados. Seu resultado foi sustentado principalmente pelo desempenho em raciocínio, de 90,7 pontos, associado aos resultados de 85,5 em linguagem e 78,7 em análise de dados. O custo estimado de US$ 0,348 por tarefa bem-sucedida também permaneceu abaixo do observado em alguns modelos proprietários com desempenho semelhante. Essa combinação sugere potencial para aplicações que demandem desempenho técnico, possibilidade de adaptação local e menor dependência de fornecedores proprietários.

O GPT-5.6 Sol Max Effort apresentou o segundo maior índice, com resultados de 91,7 em raciocínio, 87,7 em linguagem e 79,8 em análise de dados. O Gemini 3.1 Pro Preview High destacou-se pelo desempenho em seguimento de instruções, com 79,1 pontos, o maior entre os modelos classificados na faixa superior. Também apresentou o menor custo desse grupo, estimado em US$ 0,286 por tarefa bem-sucedida. Essas características podem ser relevantes para atividades que dependam da execução consistente de protocolos, da classificação de registros e do processamento padronizado de relatos ou formulários.

O GPT-5.5 Thinking xHigh Effort obteve o maior resultado em análise de dados entre os modelos destacados, com 81,6 pontos. O Claude Fable 5 Max Effort, por sua vez, apresentou o melhor desempenho em linguagem, com 90,7, além de resultados elevados em análise de dados e raciocínio. Entretanto, seu custo estimado de US$ 1,439 por tarefa bem-sucedida foi o maior do conjunto, o que pode limitar sua adoção em iniciativas com restrições orçamentárias ou grande volume de processamento.

Os modelos classificados com aderência moderada apresentaram índices entre 6,6 e 11,9. O GPT-5.4 Thinking xHigh Effort liderou esse grupo, com 11,9, seguido pelo GPT-5.6 Terra Max Effort, com 11,4, e pelo Claude 5 Opus Thinking Max Effort, com 11,0. Embora esses modelos tenham alcançado resultados elevados em raciocínio, seus índices agregados foram reduzidos por desempenhos menos equilibrados nas demais dimensões. O GPT-5.6 Terra Max Effort, por exemplo, obteve 90,6 pontos em raciocínio, mas 64,6 em seguimento de instruções.

Entre os modelos abertos classificados com aderência moderada, o DeepSeek V4 Pro apresentou índice de 9,7 e o menor custo de toda a seleção, estimado em US$ 0,050 por tarefa bem-sucedida. Embora seu desempenho técnico tenha sido inferior ao dos modelos mais bem posicionados, a combinação entre baixo custo e abertura pode torná-lo adequado para protótipos, experimentações e projetos com infraestrutura ou orçamento limitados. O GLM-5.2, com índice de 8,6, e os modelos Kimi K2.6 Thinking e Kimi K2.7 Code, com 7,4 e 6,6, respectivamente, também constituíram alternativas

abertas, ainda que apresentassem resultados inferiores, principalmente em seguimento de instruções e análise de dados.

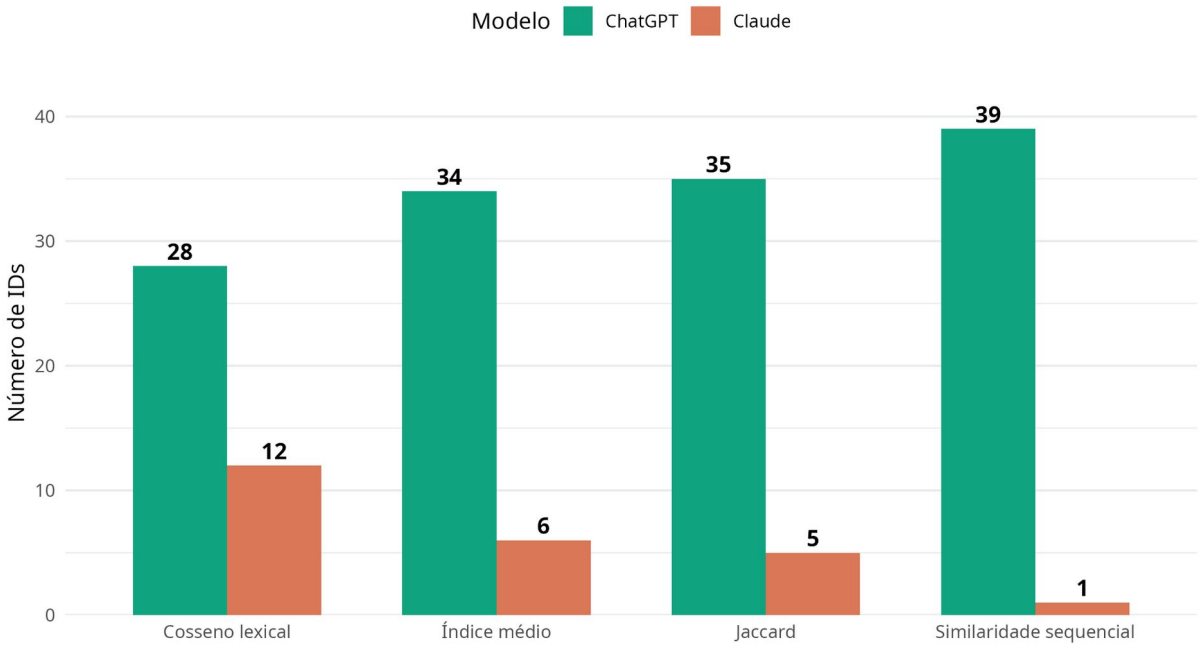
De modo geral, os resultados indicam que o desempenho técnico isolado não determina a aderência potencial às tecnologias sociais. Modelos com elevada capacidade de raciocínio podem ocupar posições inferiores quando apresentam custos elevados, menor capacidade de seguir instruções ou ausência de abertura. Em contrapartida, modelos abertos e economicamente acessíveis podem adquirir relevância em projetos que valorizem adaptação local, autonomia institucional, transparência e sustentabilidade financeira.

Apesar da ordenação obtida, os valores permaneceram baixos diante da escala nominal de 0 a 100, com resultado máximo de 13,6. O índice deve, portanto, ser interpretado como instrumento exploratório e comparativo, e não como medida absoluta de adequação. Além disso, ele não mensura diretamente impacto social, participação comunitária, acessibilidade, apropriação pelos usuários, proteção de dados ou adequação ética e contextual. A seleção definitiva dos modelos deverá ser complementada por avaliações empíricas, participativas e conduzidas nos contextos específicos de aplicação.

3.3 Resultados da experimentação

Conforme a Figura 3, os resultados indicam desempenho superior do ChatGPT em todas as métricas avaliadas. Isto confirma com os resultados apresentados na Seção 3.2 que mostrou os modelos da OpenAI superiores em relação aos modelos da Anthropic.

Figura 3 - Contagem de vitórias nas comparações



Fonte: Dados da pesquisa (2026)

No cosseno lexical, o ChatGPT apresentou a maior similaridade em 28 dos 40 IDs, correspondendo a 70% das comparações, enquanto o Claude obteve 12 vitórias, equivalentes a 30%. Embora essa tenha sido a menor diferença observada entre os modelos, o resultado demonstra maior proximidade lexical e temática das respostas do ChatGPT em relação ao gabarito.

No índice de Jaccard, a diferença tornou-se mais acentuada. O ChatGPT apresentou maior sobreposição de termos com o texto de referência em 35 IDs, ou 87,5% dos casos. O Claude foi superior em apenas cinco comparações, equivalentes a 12,5%. Esse resultado sugere que as respostas do ChatGPT compartilharam, com maior frequência, os elementos lexicais considerados relevantes nas respostas de referência.

A maior disparidade foi identificada na similaridade sequencial. O ChatGPT obteve o maior valor em 39 dos 40 IDs, representando 97,5% das comparações, enquanto o Claude foi superior em somente um caso, ou 2,5%. Esse desempenho

evidencia que as respostas do ChatGPT apresentaram maior proximidade em relação à estrutura, à ordenação e ao encadeamento textual do gabarito.

O índice médio confirmou a tendência identificada nas métricas individuais. Ao combinar os três indicadores, o ChatGPT foi superior em 34 IDs, correspondentes a 85% das comparações, enquanto o Claude apresentou maior índice médio em seis casos, equivalentes a 15%. A redução em relação às vitórias observadas nas métricas de Jaccard e similaridade sequencial pode ser explicada pelo desempenho relativamente mais equilibrado no cosseno lexical, no qual o Claude obteve sua maior quantidade de vitórias.

Em conjunto, os resultados demonstram que o ChatGPT produziu respostas mais próximas do gabarito na maior parte dos casos, sobretudo quando consideradas a sequência textual e a sobreposição vocabular. Entretanto, os indicadores devem ser interpretados como medidas de similaridade, e não como equivalentes diretos de correção conceitual. Uma resposta pode empregar palavras diferentes e preservar adequadamente o significado, assim como pode apresentar elevada proximidade lexical sem contemplar integralmente os critérios técnicos associados às tecnologias sociais. Por essa razão, as métricas quantitativas são mais adequadas quando utilizadas como instrumentos complementares a uma avaliação qualitativa ou semântica das respostas.

Declaração sobre o uso de inteligência artificial

Foi utilizada a ferramenta ChatGPT exclusivamente para revisão ortográfica, gramatical e aprimoramento da clareza textual. Para esse fim, foram empregados prompts avançados e mecanismos de controle (*guardrails*) destinados a restringir a atuação da ferramenta à correção formal do texto. Nenhuma ideia, interpretação, argumento ou conclusão gerada pela inteligência artificial foi incorporada ao conteúdo final. A responsabilidade integral pela elaboração, validação e revisão científica do texto permanece com o autor.

REFERÊNCIAS

BECKER, Katrin. Truth, text, and technology: critical reflections on the institutional implications of generative artificial intelligence. **AI & Society**, London, v. 41, n. 6, p. 5523–5530, 2026. DOI: 10.1007/s00146-026-02965-3. Disponível em: https://doi.org/10.1007/s00146-026-02965-3. Acesso em: 25 jul. 2026.

BOUCHARD, Mathieu; JOUSSELME, Anne-Laure; DORÉ, Pierre-Emmanuel. A proof for the positive definiteness of the Jaccard index matrix. **International Journal of Approximate Reasoning**, [S. l.], v. 54, n. 5, p. 615–626, 2013.

BRAUN, Robert; ZANOTTI, Laura. Challenging the cosmocracy of certainty and engaging the AI hyperpharmakon: regulation, ethics, and resonance. **International Political Sociology**, Oxford, v. 20, n. 3, art. olag023, 2026. DOI: 10.1093/ips/olag023. Disponível em: https://doi.org/10.1093/ips/olag023. Acesso em: 25 jul. 2026.

CUESTA-VALIÑO, Pedro; PENELAS-LEGUÍA, Azucena; LÓPEZ-SANZ, José María; LORANCA-VALLE, Cristina. Digital transformation, happiness-based leadership and artificial intelligence: rethinking organizational futures through happiness management. **Management Research: The Journal of the Iberoamerican Academy of Management**, [S. l.], 2026. Publicação antecipada. DOI: 10.1108/MRJIAM-07-2025-1792. Disponível em: https://doi.org/10.1108/MRJIAM-07-2025-1792. Acesso em: 25 jul. 2026.

DEL PERCIO, Alfonso. Technofuturist registers for AI and the future of work. **Journal of Sociolinguistics**, Hoboken, v. 29, n. 5, p. 319–331, 2025. DOI: 10.1111/josl.12714. Disponível em: https://doi.org/10.1111/josl.12714. Acesso em: 25 jul. 2026.

DEVLIN, Jacob et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE

TECHNOLOGIES, 2019, Minneapolis. **Proceedings [...]**. Minneapolis: Association for Computational Linguistics, 2019. p. 4171–4186. DOI: 10.18653/v1/N19-1423.

DINIS, Marisa Catarina da Conceição. Exclusively human or cyber manager? Virtus in medium est. **Revista e-Mercatoria**, Bogotá, v. 25, n. 1, p. 159–188, 2026. DOI: 10.18601/16923960.v25n1.06. Disponível em: https://doi.org/10.18601/16923960.v25n1.06. Acesso em: 25 jul. 2026.

DUTU-BUZURA, Andrei. Regulating the environmental impact of AI in European law: technology stakes and legal challenges. **Juridical Tribune: Review of Comparative and International Law**, Bucharest, v. 16, n. 2, p. 309–327, 2026. DOI: 10.62768/TBJ/2026/16/2/08. Disponível em: https://doi.org/10.62768/TBJ/2026/16/2/08. Acesso em: 25 jul. 2026.

FADILLAH, Dani; JANDEVI, Uspal. Modeling understanding through synthetic empathy and the sociotechnical reconfiguration of care. **Social Science Information**, London, 2026. Publicação antecipada. DOI: 10.1177/05390184261453006. Disponível em: https://doi.org/10.1177/05390184261453006. Acesso em: 25 jul. 2026.

FRANCESE, Rita; NAZEMI, Kawa; DESANTIS, Laura; STAFFA, Mariacarla; IASEVOLI, Felice. EMpowering PAtients THrough AI, multimedia, and explainable HCI: innovations in personalized healthcare — EMPATH-IA. In: CHITALY 2025: 16TH BIANNUAL CONFERENCE OF THE ITALIAN SIGCHI CHAPTER, 2025, Salerno. **Proceedings [...]**. New York: Association for Computing Machinery, 2025. art. 115. DOI: 10.1145/3750069.3757876. Disponível em: https://doi.org/10.1145/3750069.3757876. Acesso em: 25 jul. 2026.

GAO, Yunfan et al. Retrieval-augmented generation for large language models: a survey. **arXiv**, [S. l.], arXiv:2312.10997, 2023.

GOTO, Jameson; RAMNARAIN, Umesh. Does culture matter in AI adoption? A predictive analysis of cultural dimensions, social influence, and personal innovativeness in the modified UTAUT model. **International Journal of Educational Technology in Higher Education**, [S. l.], v. 23, n. 1, art. 13, 2026. DOI: 10.1186/s41239-026-00588-8. Disponível em: https://doi.org/10.1186/s41239-026-00588-8. Acesso em: 25 jul. 2026.

HERRERO DE LA PUENTE, Mercedes; SANCHO BELINCHÓN, Celia; SEDEÑO LÓPEZ, Jorge. IA as a tool to combat disinformation: approaching a model focused on hoaxes in an electoral context. **Doxa Comunicación**, Madrid, n. 41, p. 511–533, 2025. DOI: 10.31921/doxacom.n41a2840. Disponível em: https://doi.org/10.31921/doxacom.n41a2840. Acesso em: 25 jul. 2026.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA. **Chamada I/2026: Projeto SAVITS**. Brasília, DF: Ibict, 2026. Disponível em: https://www.gov.br/ibict/pt-br/acesso-a-informacao/oportunidades/editais-oportunidades-1/2026/chamada-i-2026-projeto-savits. Acesso em: 27 jul. 2026.

JONES, Michael N.; MEWHORT, Douglas J. K. Representing word meaning and order information in a composite holographic lexicon. **Psychological Review**, [S. l.], v. 114, n. 1, p. 1–37, 2007.

LU, Yifei. Digital divides in the AI era: socioeconomic stratification and the dual role of digital payments in China. **International Sociology**, London, v. 41, n. 3, p. 321–344, 2026. DOI: 10.1177/02685809251414400. Disponível em: https://doi.org/10.1177/02685809251414400. Acesso em: 25 jul. 2026.

NIELSEN, Aileen. Can IA, as such, invade your privacy? An experimental study of the social element of surveillance. **Indiana Law Journal**, Bloomington, v. 100, n. 4, p. 1673–1741, 2025.

RATCLIFF, John W.; METZENER, David E. Pattern matching: the Gestalt approach. **Dr. Dobb's Journal**, [S. l.], v. 13, n. 7, p. 46–51, 59, 68–72, jul. 1988.

RUVALCABA-GOMEZ, Edgar A.; GARCIA-BENITEZ, Victor H. Governance of artificial intelligence in the public sector: analysis of public policies in Spain and Mexico. **Review of Policy Research**, Hoboken, v. 43, n. 3, 2026. Publicação antecipada. DOI: 10.1111/ropr.70057. Disponível em: https://doi.org/10.1111/ropr.70057. Acesso em: 25 jul. 2026.

SEMEL, Beth M. "Everything is a signal": speaking circuits and noisy signs in the making of language-oriented AI. **Journal of the Royal Anthropological Institute**, Hoboken, 2026. Publicação antecipada. DOI: 10.1111/1467-9655.70135. Disponível em: https://doi.org/10.1111/1467-9655.70135. Acesso em: 25 jul. 2026.

SILVA, Guilherme Cavalcante. Regional integration as a remedy against technological dependency? The contradictions in the calls for regional integration in Latin American AI strategies. **Tapuya: Latin American Science, Technology and Society**, Abingdon, v. 8, n. 1, art. 2545040, 2025. DOI: 10.1080/25729861.2025.2545040. Disponível em: https://doi.org/10.1080/25729861.2025.2545040. Acesso em: 25 jul. 2026.

SUN, Hongye; GAO, Gongjing. The impact of intelligent automation on subjective well-being and job satisfaction: a comparison between standard and nonstandard employment. **China Economic Review**, Amsterdam, v. 94, pt. B, art. 102524, 2025. DOI: 10.1016/j.chieco.2025.102524. Disponível em: https://doi.org/10.1016/j.chieco.2025.102524. Acesso em: 25 jul. 2026.

TREFLER, Daniel. An economic case for regulating artificial intelligence and implications for policy. **Canadian Public Policy**, Toronto, v. 52, n. 1, p. 170–177, 2026. DOI: 10.3138/cpp.2026-001. Disponível em: https://doi.org/10.3138/cpp.2026-001. Acesso em: 25 jul. 2026.

VACHA, Juan Carlos Galindo. Artificial intelligence — AI — in the framework of digital democracy. **Revista de Derecho Político**, Madrid, n. 122, p. 159–184, 2025. DOI: 10.5944/rdp.122.2025.44745. Disponível em: https://doi.org/10.5944/rdp.122.2025.44745. Acesso em: 25 jul. 2026.

WANG, Jiayi; SANG, Guoyuan; WANG, Kai; HU, Huiqi; RUAN, Qianqian. How AI-PCK predicts AI-assisted instruction intention: a study with pre-service teachers in China. **Journal for the Study of Education and Development**, Abingdon, v. 49, n. 1, p. 95–120, 2026. DOI: 10.1177/02103702251408953. Disponível em: https://doi.org/10.1177/02103702251408953. Acesso em: 25 jul. 2026.

WANG, Ke; ZHANG, Su. AI-guided and human-guided tours in museum educational tourism: mechanisms of teenagers' long-term memory formation. **Tourism Review**, Bingley, 2026. Publicação antecipada. DOI: 10.1108/TR-09-2025-1169. Disponível em: https://doi.org/10.1108/TR-09-2025-1169. Acesso em: 25 jul. 2026.

ZAJKO, Mike. A sociology without humans: artificial intelligence as a threat to humanity, and sociology's response. **Current Sociology**, London, 2026. Publicação antecipada. DOI: 10.1177/00113921261426574. Disponível em: https://doi.org/10.1177/00113921261426574. Acesso em: 25 jul. 2026.

gov.br Documento assinado digitalmente
RAFAEL ROCHA
Data: 15/10/2026 10:59:09-0300
Verifique em https://validar.iti.gov.br

Rafael Rocha
Pesquisador

De acordo:

gov.br Documento assinado digitalmente
WASHINGTON LUIS RIBEIRO DE CARVALHO SEG
Data: 16/10/2026 17:44:04-0300
Verifique em https://validar.iti.gov.br

Washington Luís Ribeiro de Carvalho Segundo
Coordenador do projeto