Informatio
31(2), 2026, e212 ISSN: 2301-1378

DOI: 10.35643/Info.31.2.12

Artículo

Estudo cienciométrico e temático da produção sobre ética e privacidade em Modelos de Linguagem de Grande Porte de Inteligência Artificial

Scientometric and thematic study of the production on ethics and privacy on Large Language Models of Artificial Intelligence

Estudio cienciométrico y temático de la producción sobre ética y privacidad en modelos de lenguaje a gran escala de inteligencia artificial.

Leonardo Agostinelli Pereira Pinheiro de Souza[a] **Conceituação, escrita e revisão.** **ORCID: 0000-0001-5452-8060**

Ricardo César Gonçalves Sant'Ana[a] **Conceituação, escrita e revisão.**

ORCID: 0000-0003-1387-4519

[a]Universidade Estadual Paulista, Marília, São Paulo, Brasil. Correio eletrônico: leonardo.pinheiro@unesp.br, ricardo.santana@unesp.br

Resumo

Conforme se popularizou o uso de chatbots de IA, lacunas quanto a vieses, uso inadequado de dados pessoais e afins tornaram-se evidentes. Em virtude de sua importância para a sociedade e visando seu uso seguro, os modelos de linguagem de grande porte de IA, Large Language Models – LLMs, devem respeitar diretrizes éticas e de privacidade. Objetiva-se: elucidar as características da produção científica em ética e privacidade em LLMs, relativo ao montante de artigos produzidos na Computação e Ciência da Informação, analisar os temas mais relevantes e as abordagens éticas consideradas. Procedimentos metodológicos: busca nas bases internacionais, computando artigos por período de tempo e área; análise de tópicos e clusterização dos abstracts com o modelo Bertopic. Como resultados, observou-se uma tendência de crescimento no volume publicado, estando a maioria dos trabalhos na Computação, com participação tímida da CI e de artigos latino-americanos. A análise de tópicos revelou oito clusters com temas variados, destacando-se saúde e educação. As abordagens éticas privilegiam a visão tecnicista e micro-focada da computação. Conclui-se ponderando a necessidade de maior participação da CI no tema, para trazer discussões éticas mais amplas e profundas, que continuem válidas diante da evolução incessante da tecnologia.

Palavras-chave: INTELIGÊNCIA ARTIFICIAL; LLM; PRIVACIDADE; ÉTICA.

Abstract

As the use of AI chatbots has become more popular, gaps regarding biases, misuse of personal data, and the like have become evident. Due to their importance to society and aiming for their safe use, Large Language Models (LLMs) must respect ethical and privacy guidelines. The objective is to elucidate the characteristics of scientific production on ethics and privacy in LLMs, relative to the amount of articles produced in Computer Science and Information Science, and to analyze the most relevant themes and ethical approaches considered. Methodological procedures: a search in international databases, computing articles by time period and area; topic analysis and clustering of abstracts using the Bertopic model. As a result, a growing trend in the published volume was observed, with the majority of works in Computer Science, with timid participation from Information Science and Latin American papers. The topic analysis revealed eight clusters with varied themes, highlighting health and education. Ethical approaches prioritize a technocratic and micro-focused view of computing. It is concluded that greater participation from Information Science (IS) in this area is necessary to bring about broader and deeper ethical discussions that remain valid in the face of the relentless evolution of technology.

Keywords: ARTIFICIAL INTELLIGENCE; LLM; PRIVACY; ETHICS.

Resumen

Con la creciente popularidad de los chatbots de IA, se han evidenciado brechas en cuanto a sesgos, mal uso de datos personales y otros problemas. Dada su importancia para la sociedad y con el objetivo de garantizar su uso seguro, los Modelos de Lenguaje a Gran Escala (Large Language Models- LLMs) deben respetar las directrices éticas y de privacidad. El objetivo es dilucidar las características de la producción científica sobre ética y privacidad en los LLMs, en relación con la cantidad de artículos publicados en Ciencia de la Computación y Ciencia de la Información, y analizar los temas y enfoques éticos más relevantes. Procedimientos metodológicos: búsqueda en bases de datos internacionales, recopilación de artículos por período y área; análisis temático y agrupamiento de resúmenes mediante el modelo Bertopic. Como resultado, se observó una tendencia creciente en el volumen publicado, con la mayoría de los trabajos en Ciencias de la Computación y una participación reducida de artículos de Ciencia de la Información y Latinoamérica. El análisis temático reveló ocho grupos con temas variados, destacando la salud y la educación. Los enfoques éticos priorizan una visión tecnocrática y microenfocada de la Computación. Se concluye que una mayor participación de la Ciencia de la Información (CI) en este ámbito es necesaria para propiciar debates éticos más amplios y profundos que sigan siendo válidos ante la constante evolución de la tecnología.

Palabras clave: INTELIGENCIA ARTIFICIAL; LLM; PRIVACIDAD; ÉTICA.

Fecha de recibido: 19/05/2026
Fecha de aceptado: 29/07/2026

1.Introdução

Pode-se traçar a gênese da Inteligência Artificial (IA) desde a Segunda Guerra Mundial, enquanto o matemático Alan Turing desenvolvia técnicas de descriptografia de mensagens, levando a estudos posteriores visando a criação de mecanismos 'inteligentes', que pudessem entender e interagir com o contexto circundante (Barcauí, 2025). Tal ideia pode ser também entendida no âmbito do auge da ciência da Cibernética, no mesmo período, criada por Nobert Wiener, definida como "o domínio todo da teoria da comunicação e do controle, seja na máquina ou no animal" (Martinelli, 2012, p. 27). Daí decorrem analogias entre as máquinas, sistemas artificiais, e os seres vivos, sistemas biológicos, tornando-os equiparáveis em seu funcionamento em termos de homeostase, auto-regulação, aprendizagem, adaptação e afins; a tal ponto que a pele humana pode ser entendida com um sensor, seu cérebro um computador complexo (Op. cit.). Oficialmente, entretanto, o nascimento da IA como campo científico remonta ao ano de 1956, lançada a agenda de pesquisa por John Maccarthy e associados (Barcaui, 2025).

Embora dotar as máquinas de qualidades antropomórficas possa parecer, possivelmente, um conceito estranho às ciências humanas e sociais, apura-se que este entendimento persiste no desenvolvimento da IA atual, tanto em seus conceitos, nomenclaturas ('neurônio artificial' e afins), quanto em sua ambição do desenvolvimento de IA dita 'forte' (ou *Artificial General Intelligence*, Inteligência Artificial Geral - AGI), capaz de desempenhar uma grande variedade de tarefas "em nível humano", ou mesmo de desenvolver consciência (Russel, Norvig, 2022). De fato, o desenvolvimento da IA com habilidades e raciocínio em nível humano foi colocado como objetivo futuro factível por empresas de tecnologia como a Meta e outras (Labarte, 2025; CORREIO BRASILIENSE, 2026).

Contudo, o estado da arte atual permitiu apenas o desenvolvimento de aplicações no paradigma denominado como IA fraca ou estreita, com capacidade para desempenhar um conjunto pequeno de tarefas especializadas, sem capacidade real

de raciocínio (Barcaui, 2025). Dentre tais capacidades da IA estreita encontra-se o processamento de linguagem natural (Natural Language Processing – NLP).

A IA generativa, antes restrita praticamente ao contexto acadêmico e certas aplicações empresariais e industriais, passou a popularizar-se entre o público leigo por meio de chatbots que utilizam NLP. Tais aplicações, como DeepSeek, Gemini, Grok, Chat GPT e afins, se baseiam na tecnologia dos grandes modelos de linguagem (Large Language Models - LLMs), que incorporam grandes volumes de dados disponíveis na Internet e fontes variadas (*big data*), para serem treinados e formarem bilhões de parâmetros que servirão para a geração de respostas.

Entretanto, para ser utilizada visando auxiliar tanto organizações quanto os sujeitos em suas tarefas cotidianas, a IA deve se mostrar confiável, respeitando as diretrizes éticas e de privacidade. Uma investigação feita pelo Washington Post revelou conteúdos de misoginia, xenofobia e similares em um conjunto de dados de treinamento do modelo GPT, o que tem o potencial para geração de respostas com vieses para tais tendências (Schau; Chen; Tiku, 2023). Ademais, tais conjuntos de dados podem conter informação pessoal sensível, o que os torna um risco potencial para a privacidade (Olff, 2025).

Em vista do panorama acima, e da complexa relação dos sujeitos e organizações com as Tecnologias da Informação e Comunicação (TICs), que representam novas possibilidades positivas, mas também novas ameaças, emergiu a ética da informação, relacionada à Filosofia da Informação. Duarte (2022) ressalta que não há uma vertente única da ética da informação, ainda que todas estas percepções se coadunem em discutir os impactos morais das TICs no cotidiano. Um exemplo seria a ética da informação de Luciano Floridi, que a trabalha não sob o olhar de contextos particulares, mas como uma teoria macro, que contempla a informação na tensão e na complementariedade entre a vida '*online'* e '*offline'*, encampando a epistemologia, metafísica, lógica e afins, trazendo ao debate valores filosóficos atemporais (Op. cit.). Outro aspecto fundamental da ética '*floridiana*' seria o bem-estar da *infosfera*, compreendida esta última como o ambiente de entrelaçamento de todos os agentes informacionais, seus processos, suas propriedades e relações (Pellegrini; Vitorino, 2018).

Assim, são propostos os seguintes problemas de pesquisa: quais são as características principais do corpus de publicações no tema ética e privacidade em

LLMs em termos quantitativos? Em termos qualitativos, quais tópicos de discussão se mostram mais relevantes nesse tema? Que abordagem ética predomina?

Como objetivo geral, é colocado: elucidar as características da produção científica em ética e privacidade em LLMs, com respeito ao montante de artigos produzidos e suas temáticas. Como objetivos específicos, são colocados: efetuar busca sobre o tema proposto em bases de dados de relevância internacional; computar a evolução da produção por período de tempo e nas áreas correlatas à Ciência da Informação e computação; agrupar os trabalhos por similaridade, extrair as palavras mais representativas em cada grupo; analisar os artigos mais relevantes em relação ao seu tema, abordagem e tipo de solução proposta; verificar a perspectiva de ética predominante nos artigos analisados.

2. Referencial teórico

É relevante conceituar as bases de alguns aspectos do aprendizado de máquina relevantes para o presente estudo, quais sejam: *deep learning* (aprendizado profundo); redes neurais *transformers*; *Natural Language Processing* (Processamento de Linguagem Natural - NLP) e *Large Language Models* (Grandes Modelos de Linguagem - LLMs). Partindo-se do princípio que os *chatbots* populares de IA atuais são fundamentados em redes neurais artificiais, que emulam o funcionamento dos neurônios biológicos do cérebro humano, o *deep learning* se baseia em redes multicamadas densamente conectadas. Tais estruturas intrincadas propiciam a execução de processamentos complexos que permitem à rede realizar tarefas normalmente atribuídas a seres humanos, tais como: reconhecimento de imagens, geração de imagens estáticas, vídeos, música, geração de textos em linguagem humana e assim por diante (Barcaui, 2025).

A densidade dessas redes, contudo, traz o efeito colateral da ‘opacidade’, que dificulta auditar como se dão os fluxos de dados e transformações na passagem da informação de um neurônio a outro, o que pode gerar certo nível de inexatidão/imprevisibilidade nos resultados gerados.

Maior acurácia e estabilidade poderia ser atingida em modelos como o *random forest*, compostos por redes de árvores de decisão, onde se pode traçar com precisão

os caminhos e transformações dos dados alimentados na rede até sua saída, malgrado seu alto custo computacional, que torna seu uso impraticável em tarefas sofisticadas de *deep learning* (Faceli et al., 2025). De maneira mais específica, o processamento de linguagem natural, anteriormente, foi realizado mediante redes neurais recorrentes (*Recurrent Neural Networks - RNNs*) e redes do tipo *Long-Short Term Memory* (LSTMs), sendo, no presente executado por redes transformers, que aprimoraram o processamento de texto, com seu mecanismo de auto-atenção (Russel; Norwig, 2022).

No mecanismo de auto-atenção cada palavra consegue guardar informação contextual de si mesma, enquanto 'olha' também para todas as demais, mantendo o nexo de relação das palavras mesmo em frases muito longas (Vaswani, 2017). Ademais, as mesmas têm a vantagem de operarem com processamento paralelo em vez de sequencial, o que representa um ganho de desempenho. Desta forma, cada um dos mecanismos de atenção paralelos, ou cabeças de atenção, consegue captar diferentes camadas de significado em um texto, como a relação sujeito-verbo, dependências de longa distância entre termos de uma frase e assim por diante.

Conforme a IBM (2025) os LLMs são redes neurais treinadas com grandes quantidades de dados, gerando bilhões de parâmetros, de modo a reconhecer padrões complexos, que os permitam gerar textos como seres humanos, dentre outras tarefas, reconhecendo contextos para gerar respostas coerentes. A curadoria dos dados fornecidos é crucial para a integridade e até mesmo para aspectos éticos dos resultados obtidos.

Como já mencionado na presente seção, a opacidade de modelos de IA generativa é preocupante no sentido de que pode haver a ingestão, intencional ou não, de dados sensíveis para sujeitos e organizações, como dados pessoais, informações de mídias sociais e bancos de dados vazados (Tong et al.). Tais dados, podem ser eventualmente expostos como saída em algum prompt de usuário, dependendo de suas intenções de busca.

Dois riscos que se corre, explicam Sant'Ana e Martins (2024), referente à abordagem de privacidade de Solove, são a agregação de dados esparsos, de tal forma que suas características combinadas possam ser suficientes para identificar o sujeito, expondo detalhes indesejados e/ou não permitidos. Dados pessoais sensíveis podem ser assim entendidos: "[...] exemplos de informações sensíveis

incluem informações relacionadas à esfera mais íntima do titular dos dados, como opiniões políticas, vida sexual ou condenações criminais" (NIST, 2024, p. 4, tradução nossa).

Destaca-se o Ciclo de Vida de Dados (CVD) como um modelo relevante para investigar os LLMs pela ótica da Ciência da Informação. Tal *frame* analisa as fases de acesso a dados, a saber: coleta, armazenamento, recuperação e descarte; levando em conta fatores transversais ubíquos em todas as fases: privacidade, integração, qualidade, legislação, disseminação e preservação (Sant'Ana, 2016). No presente artigo dá-se enfoque no fator privacidade e suas implicações éticas.

Concernente à privacidade da informação, é preciso entender que aspectos culturais podem influenciar a percepção do conceito de privacidade aceito por uma sociedade. Capurro (2006) menciona que a privacidade da informação no Ocidente enfoca o aspecto da subjetividade, enquanto que o olhar de culturas asiáticas e africanas é mais prático e focado em valores coletivos. De acordo com este autor (2006), a perspectiva tradicional budista enxerga a individualidade como algo negativo, por esta razão, o entendimento dos japoneses sobre a privacidade no ciberespaço repousa sobre uma dicotomia onde: por um lado os valores ocidentais parcialmente assimilados clamam pelo direito de o sujeito controlar suas próprias informações; por outro, a web é encarada como um espaço público onde o indivíduo tem relevância secundária.

Uma das técnicas utilizadas para extrair dados sigilosos eventualmente utilizados no treinamento de IAs são os chamados Membership Inference Attacks (ataques de inferência de membro - MIA). Nesta modalidade, o atacante consulta o LLM repetidamente para verificar, via testes estatísticos com amostras, se uma informação sensível consta em seu conjunto de dados de treinamento (IBM, 2025). Conforme o nível de acesso que os atacantes tenham sobre os dados de treinamento da rede, pode-se classificá-los, do maior nível de acesso até a total falta de acesso, como ataques de caixa branca, cinza e preta. Na maior parte das vezes, os atacantes não terão qualquer acesso aos dados e configurações, necessitando utilizar técnicas avançadas de inferência de caixa preta, que têm evoluído rapidamente e se tornado bastante eficazes (Shi et al., 2025).

Algumas estratégias, no entanto, podem ser utilizadas para mitigar o impacto de dados pessoais e/ou sigilosos no treinamento e geração de resultados da IA, como,

por exemplo, a privacidade diferencial (Differential Privacy - DP). Esta técnica faz com que, havendo dois ou mais modelos semelhantes, que o impacto da existência de um conjunto de dados único e distinto entre os modelos seja estatisticamente insignificante, indistinguível; isto se dá por meio da limitação de impacto de dados discrepantes (*outliers*), adição proposital de ruídos, dentre outras técnicas (Ye et al., 2022; Rigaki; Garcia, 2023). Esta abordagem, contudo, é computacionalmente custosa e pode afetar o desempenho do modelo.

Um aspecto de suma importância nos modelos de IA generativa baseados em texto é referente a implantar medidas de segurança que visem impedir a geração de conteúdo pernicioso e/ou ilegal. Em casos em que o usuário requisite conteúdos proibidos a IA geralmente deve se negar a responder. Contudo, novos meios de burlar tais salvaguardas (*guardrails*) têm sido desenvolvidos, principalmente por equipes de pesquisadores que buscam explorar brechas de segurança em tais sistemas. Dentre as técnicas para burlar dispositivos de segurança de IAs encontra-se o uso de sufixos adversários (*adversarial suffix*). A técnica desenvolvida por Zou et al (2023) consiste em adicionar um *token*, ou cadeia de caracteres especial após o prompt ou frase de entrada do usuário com a finalidade de burlar os mecanismos de segurança do modelo. O método utiliza gradientes (Greedy Coordinate Gradients) para gerar automaticamente *tokens* que tenham maior probabilidade de obter uma resposta positiva do modelo. O método verifica que *tokens* teriam menor gradiente de perda, ou maior chance de acerto realizando e testando várias combinações desses *tokens* (Zou et al., 2023).

Outro tipo de ataque é o *role play* (interpretação de papel). Em tal ação, o perpetrador solicita à IA simular ser um agente mal intencionado, ‘enganando-a’ para que baixe a guarda de seus mecanismos de segurança (Shanahan et al., 2023). Outros ataques com objetivos similares são: a injeção de *prompts*, onde o perpetrador insere código malicioso em seu texto para que o LLM ignore as instruções dos programadores e atenda a solicitações ilícitas (dentre essas técnicas a *adversarial suffix*); técnica de múltiplas tentativas, que manipula o comportamento da IA com sequências de *prompts* encadeados ao longo do tempo; *many-shot*, que sobrecarrega a IA com muitas requisições em um mesmo *prompt*, explorando seu limite de janela de contexto, tentando enfraquecer suas salvaguardas (Krantz; Jonker, 2025). O envenenamento de dados é também um tipo de ataque

bastante danoso, que age inserindo dados maliciosos no corpus de treinamento dos modelos de LLM, visando prejudicar seu desempenho geral e/ou deixar brechas de segurança que possam ser exploradas posteriormente (Zhao et al., 2025).

Subjacente ao fator privacidade, está o aspecto ético, sendo os dois conceitos inextricáveis. Pode-se classificar os estudos éticos nas TICs em basicamente dois paradigmas: um micro-focado, que discute contextos de problemas particulares nesse âmbito, e um macro-focado, que enxerga as TICs pela ótica de aspectos filosóficos mais abrangentes e aprofundados. No primeiro paradigma, afirma Gonzalez (2013), está a ética da computação, seguindo dois caminhos: o de fomentar uma conscientização nos profissionais e na sociedade sobre os impactos tecnológicos; buscar estratégias de contingência para remediar os possíveis impactos negativos das mesmas tecnologias. Nesta linha segue o desenvolvimento das atuais políticas regulatórias de IA em vários países, com um enfoque jurídico e técnico, com uma visão mais pragmática (Feferbaum et al., 2023).

Em contraste com a visão utilitarista acima descrita, está a filosofia da informação, principalmente sob a perspectiva '*floridiana*'. Esta olha para a informação desde sua gênese e seus impactos positivos e negativos no entorno social de modo amplo, considerando que a abordagem micro-focada resulta em soluções parciais e limitadas (Gonzalez, 2013).

Ademais, Pellegrini e Vitorino (2018) realizam uma síntese do conceito de ética informacional mediante a análise de vários autores da Ciência da Informação, afirmando que esta diz respeito às ações humanas, orientando-as para boa conduta, num contexto de dilemas e distorções comunicacionais e informacionais, onde é preciso equilibrar valores conflitantes, tendo como objetivos: uma 'vida boa', a justiça e o bem coletivo. Embora este entendimento da ética informacional seja aplicado originalmente à competência em informação, o mesmo se mostra suficientemente amplo para atender aos propósitos da presente pesquisa.

Outra perspectiva que se coaduna à apresentada acima é a exposta por Capurro (2008), onde a sociedade da informação deve presar por valores como a liberdade, igualdade, solidariedade, tolerância e partilha de responsabilidade; enquanto que a ética da informação deve presar pela justiça, dignidade e valorização da pessoa humana. Ainda que a ética da informação aspire à universalidade em seus valores, sua aplicação pode estar sujeita às perspectivas distintivas de cada sociedade, como

no caso da africana, onde se alia ao conceito ancestral de ubuntu, que representa o ímpeto de cuidar e compartilhar (Capurro, 2008). Esta constatação é coerente com o fato anteriormente apresentado das distinções culturais quanto à privacidade da informação.

Outro aspecto ético de suma relevância concernente aos chatbots de IA é o nível em que podem interferir na percepção da realidade mediante as informações por eles providas. Capurro (2020) argumenta que os algorítimos computacionais, enquanto tecnologias, além de serem moldados por seres humanos podem também moldar a mentalidade destes e seu modo de ser. Capurro argumenta que tais algorítimos não devem ser considerados como entidades literalmente autônomas, visto não serem capazes de julgamento moral de suas ações, o que exige uma regulação humana que leve em conta fatores contextuais e culturais no emprego dessas tecnologias.

Em virtude da constatação da incapacidade moral e volitiva do algoritmo computacional e da IA, e da existência do valor ético fundamental da responsabilização já discutido, sugere-se que a responsabilidade moral deve ser assumida pelos desenvolvedores das tecnologias, principalmente. Entretanto, ao usuário caberia a responsabilidade de salvaguardar seus dados pessoais sensíveis, evitando exposições desnecessárias, sempre e quando lhe for dada esta opção. Tal estratégia torna-se vital, principalmente em tempos de uso intensivo de redes sociais. Este é precisamente o cerne da técnica de obfuscação proposta por Capurro (2020), que visa driblar o excessivo controle das TICs sobre os dados pessoais sensíveis.

Por fim, salienta-se a necessidade de uma visão sociotécnica consistente sobre a IA, que não apenas vise mitigar os problemas presentes, mas que seja suficientemente abrangente e profunda para compreender e problematizar a influência dessa tecnologia na sociedade no longo prazo.

3. Procedimentos metodológicos

A presente pesquisa se trata de um estudo cienciométrico com uma abordagem qualiquantitativa. Inicialmente, foi realizada uma busca nas bases Scopus, Web of

Science (WoS) e Scielo sobre os temas LLM, ética e privacidade, filtrando pelo título, palavras-chave e abstract. Foram considerados artigos em inglês e português, de áreas correlatas à computação e Ciência da Informação, ou Ciências Sociais, quando a área da Ciência da Informação não estava disponível na base. O período de tempo foi determinado desde 2017, ano em que o trabalho seminal de Vaswani et al. (2017) lançou a tecnologia de redes *transformer*, essencial para o ganho de desempenho e popularização dos LLMs e chatbots de IA atuais. O protocolo completo de busca pode ser observado no Quadro 1, contendo os critérios de inclusão e exclusão de trabalhos.

Quadro 1: Protocolo de busca

Estratégias de busca	
Data da busca	Março/2026
Scopus	TITLE-ABS-KEY ("Large Language Model*" OR LLM) AND TITLE-ABS-KEY ("privacy" OR "ethic*") AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (OA , "all")) AND (LIMIT-TO (LANGUAGE , "English") OR LIMIT-TO (LANGUAGE , "Portuguese")) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "SOCI")) AND (LIMIT-TO (PUBSTAGE , "final"))
WoS	TS= ("Large Language Model*" OR LLM) AND TS= ("privacy" OR "ethic*")
Scielo	("Large Language Model*" OR LLM) AND (privacy OR ethic*)
Campos	Título, resumo e palavras-chave
Período	2017 - 2026
Critérios de inclusão de trabalhos	
Tipo doc.	Artigos de periódico
Status	Final
Línguas	Inglês e português
Áreas	Ciência da Computação e Ciência da Informação ou Ciências Sociais
Critérios de exclusão de trabalhos	
Duplicação	Ocorrência do mesmo trabalho em diferentes bases de dados.
Tipo doc.	Artigos de anais de eventos, capítulos de livro e afins.
Status	No prelo
Línguas	Quaisquer que não português ou inglês.
Período	Artigos publicados antes de 2017.
Áreas	Não relacionadas à Computação Ciência da Informação, ou Ciências Sociais.

Fonte: Elaborado pelos autores (2026).

Numa primeira etapa, utilizou-se a linguagem R para analisar quantitativamente as características do corpus de artigos recuperados, utilizando-se principalmente de aportes da esteatítica descritiva, formando também tabelas e gráficos para visualização dos dados.

Na etapa qualitativa elaborou-se um código em Python para gerar agrupamentos por similaridade de assunto dos artigos. Inicialmente, os resultados de busca nas bases foi baixado no formato CSV, sendo realizada a exclusão de todos os artigos duplicados, identificados pela duplicidade de seus títulos.

Os *abstracts* dos artigos foram convertidos em *embeddings* com o modelo Sentence Bert (Sbert). Os *embeddings* são vetores numéricos de alta dimensionalidade, representando a proximidade semântica das palavras presentes. O modelo Bertopic foi então alimentado com os *embeddings* para a geração de tópicos de assuntos e formação de *clusters* por afinidade semântica entre os *abstracts*. O Bertopic é uma biblioteca Python para modelagem de tópicos com funções robustas de NLP utilizando redes transformers (Bertopic, 2026).

Em seguida, foram recuperados os títulos dos artigos mais representativos de cada *cluster*, segundo o cálculo de probabilidade do Bertopic. Tais artigos foram então lidos na íntegra e analisados qualitativamente, para determinar seu tema principal, abordagem metodológica e tipo de solução proposta.

4.Resultados e discussões

Sendo que os LLMs popularizaram a IA entre o público geral, leigo, por meio dos *chatbots*, é pertinente examinar como o seu advento impactou também no âmbito acadêmico. Relevante também é tratar dos dilemas éticos e de privacidade que emergiram conforme aspectos como vieses, vazamentos de dados pessoais, alucinações e outras ocorrências passaram a se tornar mais perceptíveis conforme os usuários interagiam com as plataformas de IA. A tabela 1, abaixo, mostra a evolução temporal da produção sobre ética e privacidade mediante busca realizada nas bases Scopus, WoS e Scielo. Salienta-se que a busca nas bases de dados foi executada em março de 2026.

Tabela 1: Artigos publicados por ano

Período	Scopus	WoS	Scielo
março/2026 (parcial)	117	43	0
2025	562 (+192,71%)	309 (+149,19%)	0
2024	192 (+242,86%)	124 (+396%)	02
2023	56 (+5500%)	25 (+2400%)	0
2022	01 (-50%)	01	0
2021	02	0	0
Total	**930**	**502**	**02**
Participação %	**64,85%**	**35,01%**	**0.14%**
Média	**155**	**83.67**	**0.33**
D. P.	**212.31**	**119.44**	**0.82**
C. V.	**136.98**	**142.76**	**244.95**

Fonte: Elaborado pelos autores (2026).

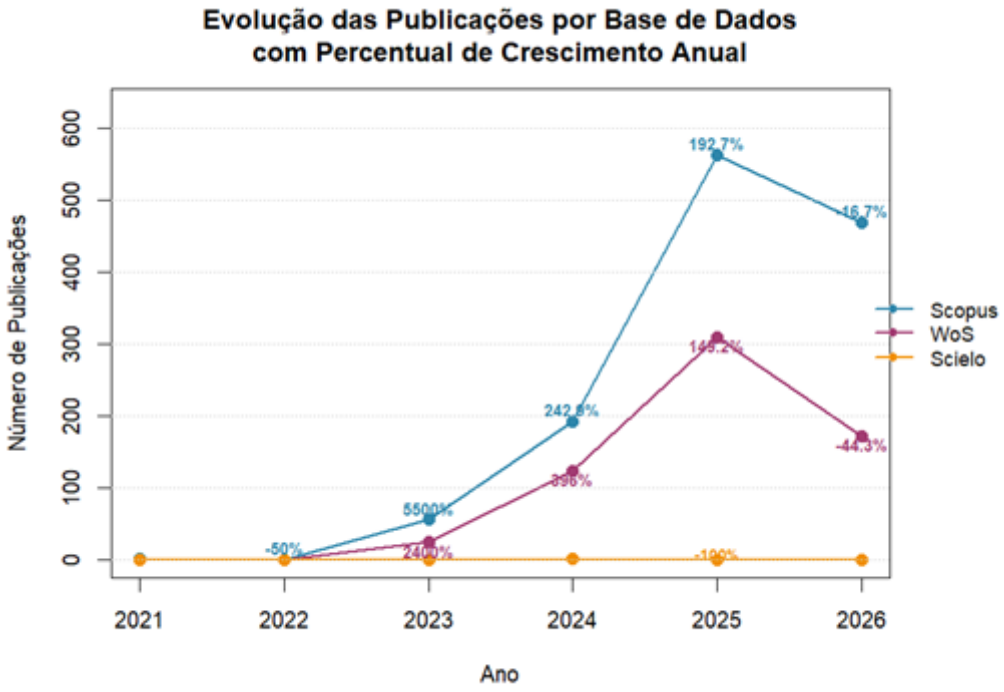
A partir de 2023, iniciou-se um padrão de aceleração na produção acadêmica nas bases Scopus, com crescimento de 5500% relativo ao ano anterior, e na base WoS, com crescimento de 2400%. Este pico de produção, possivelmente, reflete o lançamento do *chatbot* popular pioneiro, Chat GPT, em novembro de 2022. A partir de então, pôde-se verificar uma popularização da IA entre as pessoas em geral, as empresas, bem como intensa cobertura midiática do assunto. É interessante observar como esse interesse popular também influenciou o campo acadêmico, levando-se em conta que a academia já estuda a IA há, pelo menos, sete décadas, como discutido anteriormente (seção 1).

Outro ponto de interesse, e preocupação, é o fato de a base Scielo não ter registrado nenhum trabalho até o ano de 2024, apresentando apenas 02 trabalhos nesse mesmo ano, e seguindo sem produção pelo restante do período pesquisado. Sua participação no montante de artigos recuperados foi de apenas 0,14%. Esta observação evidencia uma necessidade premente de que os autores latino-americanos acompanhem o restante do mundo na investigação deste tema de suma relevância para o contexto hodierno, bem como prestigiem seus periódicos nacionais, fomentando o acesso e o debate sobre o tema no sul global. Observa-se, a partir de 2024 uma continuidade no crescimento da produção na Scopus e na Wos, contudo em um ritmo menos acelerado. Salienta-se que, visto os dados terem sido coletados no início de março de 2026, os dados desse ano são parciais.

Interessante também é analisar o desvio-padrão (D. P.) da produção em cada base de dados. Na base Scopus o D. P. se mostrou alto, 212,31, indicando grande

oscilação da produção ano a ano, partindo de valores bastante baixos para um pico de produção em um curto período. O D. P. representado pela base WoS foi também alto, 119,44, embora de amplitude menor que na Scopus. Por fim, o D. P. na Scielo, 0,82, parece também alto, mas isto se deve à sua sequência de valores ser composta quase toda de zeros, com exceção de 2024, onde obteve-se 02 publicações. Igualmente, o Coeficiente de Variação (C.V.) também se mostra muito alto em todas as bases, mostrando que os números variam drasticamente. Estes resultados refletem, possivelmente, um tema de pesquisa em franco crescimento, sendo este crescimento um fenômeno recente, uma tendência de estudos emergente. A figura 1, abaixo, traz uma dimensão visual do comportamento da produção acadêmica sobre o tema pesquisado ao longo do tempo.

Figura 1: Representação da produção por ano



Fonte: Elaborado pelos autores (2026).

Embora os dados tenham sido coletados em março de 2026, é possível fazer uma previsão do montante de artigos que serão publicados até o fim desse ano, levando-se em consideração a média mensal multiplicada por 12 meses. Observa-se, na Tabela 2, um decréscimo possível na produção, com estagnação da base Scielo,

coerente com a desaceleração já observada em 2025, embora a tendência geral seja de crescimento, como mostraram as análises do D.P. e C.V..

Tabela 2: Projeção de produção para 2026.

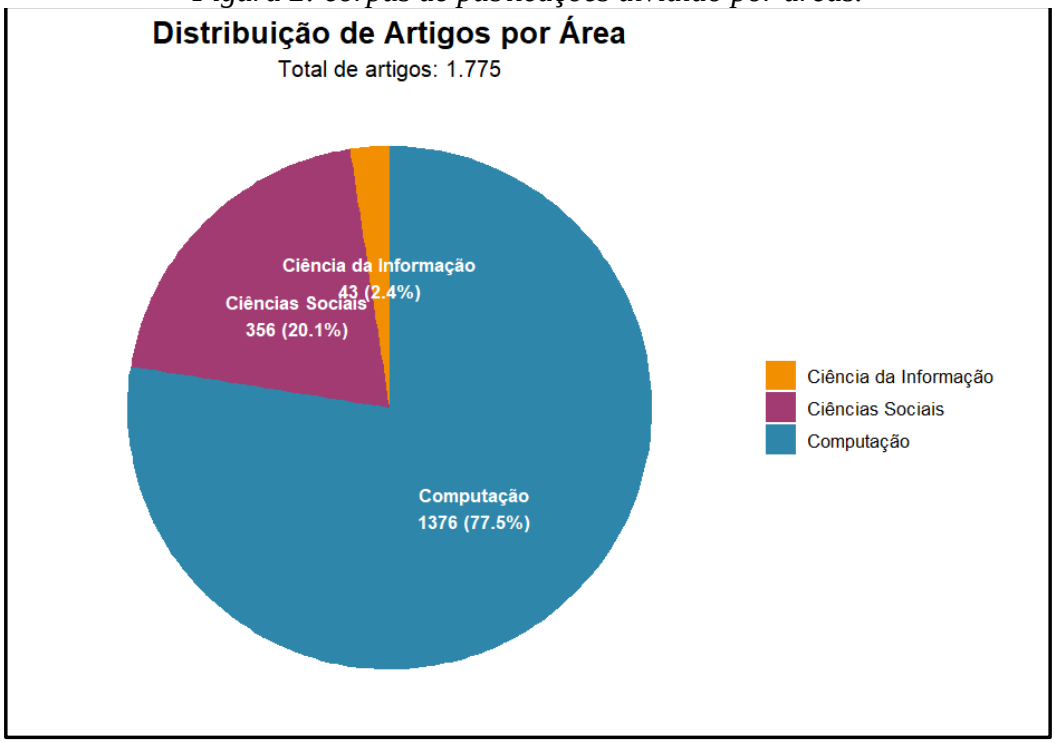
	Qtd. março/2026	**Média/mês 2026**	**Previsto dez/2026**
Scopus	117	39	468 (-16,73%)
WoS	43	14,33	172 (-44,34%)
Scielo	0	0	0

Fonte: Elaborado pelos autores (2026).

Concernente à projeção realizada para o ano de 2026, justifica-se sua relevância para verificar o comportamento do volume publicado no futuro próximo e constatar a tendência de crescimento, visto que o tema de pesquisa despertou maior interesse apenas muito recentemente, a partir de 2023, como mostram os dados coletados. A projeção, portanto, surge como estratégia válida para amplificar a visão sobre o tema e estender o campo de análise num horizonte temporal que se mostra reduzido, tomando-se apenas os dados concretos.

Na Figura 2, abaixo, pode-se verificar a distribuição do total do corpus de artigos recuperados por área do conhecimento. Ressalta-se que o filtro de área da Ciência da Informação não está disponível nas bases Scielo e Scopus, apenas na WoS. Por este motivo, nestas bases, foi selecionada a área de Ciências Sociais em seu lugar, visto que é a grande área que contempla a Ciência da Informação. Contudo, constatou-se que a Ciência da Informação em si perfaz a pequena fatia de 2,4% do corpus de artigos, seguida por 20,1% das Ciências Sociais. A vasta maioria dos artigos recuperados é da área da Computação, 77,5%.

Figura 2: corpus de publicações dividido por áreas.



Fonte: Elaborado pelos autores (2026).

Os dados acima apresentados são em certa medida preocupantes, pois demonstram que as duas áreas que se empenham pela criticidade na análise da relação da informação no contexto social e humanístico desempenham, aparentemente, um papel marginal neste debate sobre ética e privacidade na IA, que parece dominado pela computação e sua visão tradicionalmente tecnicista, utilitarista. Findada a análise quantitativa, passa-se à parte qualitativa da pesquisa.

Para iniciar a etapa qualitativa da pesquisa, todos os três arquivos CSV com dados dos artigos recuperados foram unificados e os trabalhos duplicados, com base na comparação de títulos, foram excluídos, restando 1061 artigos únicos.

A seguir, um código em Python foi desenvolvido, utilizando o modelo Sbert para converter os *abstracts* em *embeddings*. Os *embeddings* foram estatisticamente comparados entre si com o modelo Bertopic para geração de tópicos de assunto. Foram formados oito tópicos e mais um conjunto de *outliers* (tópico -1), que representa os trabalhos com nível de semelhança muito baixo para se integrarem em algum tópico ou *cluster*. A este conjunto de dados discrepantes, *outliers*, se denomina ruído. A Tabela 3 mostra as quantidades de artigos por tópico e algumas análises estatísticas.

Tabela 3: Quantidades de artigos por cluster

Tópicos	**Artigos**	**%**
Tópico -1 (ruído- *outliers*)	247	23,28%
Tópico 0	214	20,17%
Tópico 1	189	17,81%
Tópico 2	146	13,76%
Tópico 3	137	12,91%
Tópico 4	51	4,81%
Tópico 5	45	4,24%
Tópico 6	16	1,51%
Tópico 7	16	1,51%
Totais	**1061**	**100%**
D. P. (sem ruído)	79,2	
C. V. (sem ruído)	77,89%	
Taxa de ruído	23,28%	
Taxa aproveitamento	76,72%	

Fonte: Elaborado pelos autores (2026).

A taxa de aproveitamento (T. A.) se refere à razão entre o total de artigos recuperados e agrupados em temas/*clusters* e os considerados como *outliers*/ruído. A T.A. calculada foi de 76,72%, demonstrando que o algoritmo Bertopic conseguiu agrupar satisfatoriamente a maioria dos artigos baseado em similaridade semântica. A taxa de ruído, portanto, se mostrou baixa: 23,28%.

Excluindo-se o grupo do ruído, *outliers*, verificou-se um desvio-padrão (D.P.) relativamente alto, de 79,2, próximo da média de artigos/tópico (101,75), o que representa uma dispersão significativa na quantidade de artigos entre os tópicos. Evidencia-se uma heterogeneidade temática e concentração em alguns tópicos dominantes (0 a 3), sendo os tópicos 6 e 7 periféricos, com o menor número de artigos. Tal variabilidade é confirmada pelo coeficiente de variação (C. V.), de 77,89%, evidenciando a concentração temática em poucos tópicos e baixa uniformidade entre os mesmos.

A seguir, o quadro 2 exibe, em ordem decrescente de tamanho, os tópicos e suas constituições, em termos das palavras mais relevantes encontradas nos abstracts, identificadas pelo Bertopic. Para cada tópico foi definido um tema baseado nestas palavras. Além dos aspectos de privacidade, ética e segurança da informação, aplicações em saúde e educação também se destacam. A ética e a privacidade na prática médica são aspectos fundamentais, assim como o uso de IA em ambientes

educacionais passou a ser uma preocupação, relativo ao fator plágio e interferências no processo de aprendizagem.

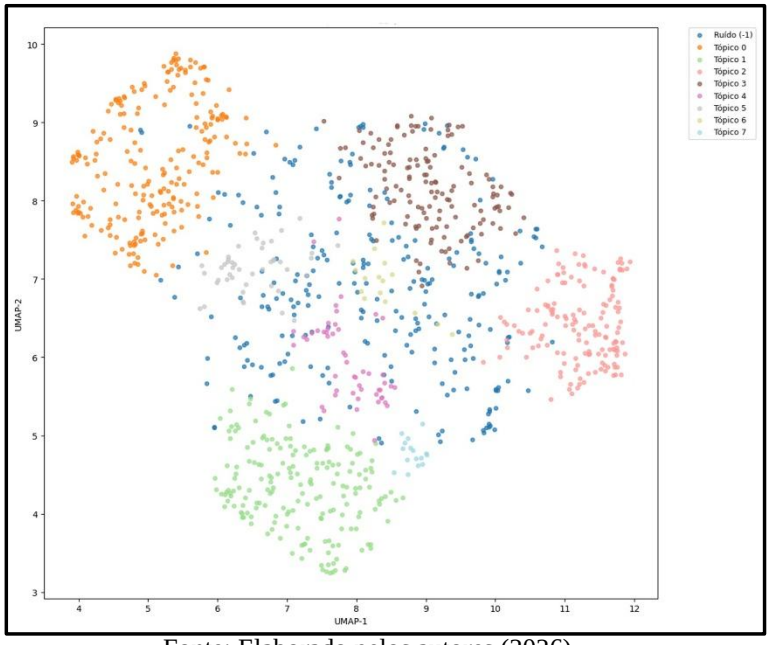
Quadro 2: tópicos gerados pelo Bertopic e suas palavras representativas.

Tópico	**Termos representativos**	**Tema**
-1	Não consta (ruído)	Não consta (ruído)
0	privacy, data, security, framework, detection, systems, information, learning, analysis, sensitive	Privacidade e segurança em IA
1	medical, healthcare, data, patient, performance, ethical, health	Ética em saúde em geral no contexto da IA
2	education, students, learning, use, academic, educational, gen ai, ethical, tools, generative	Aspectos sociotécnicos da IA na educação
3	ethical, bias, human, moral, artificial, systems, development, potential, alignment	Valores éticos na IA em geral
4	health, support, mental, user, mental health, interaction, data, care, healthcare, users	Aplicações da IA em saúde mental
5	data, rag, generation, retrieval augmented, knowledge, retrieval, retrieval augmented generation, framework, code, tasks	Representação e recuperação do conhecimento com IA
6	text, tools, misinformation, influence, generated, narratives, ai generated, information, simulated, social	Idoneidade da informação gerada por IA e seu impacto social
7	use, health, care, health care, perceived, trust, concerns, chatbots, participants, chatbot	Aplicações de IA em saúde focadas no usuário

Fonte: Elaborado pelos autores (2026).

Na figura 3, abaixo, é exibida uma representação gráfica da distribuição dos clusters de tópicos.

Figura 3: Representação visual dos clusters/tópicos gerados.



Fonte: Elaborado pelos autores (2026).

Por fim, no quadro 3 é mostrado um compilado de todos os *clusters* gerados, das palavras representativas de cada *cluster*/tópico, bem como uma análise dos artigos mais relevantes de cada *cluster*. A relevância dos artigos foi calculada automaticamente pelo Bertopic baseado no maior índice de semelhança do perfil de palavras do *abstract* do artigo com o perfil de palavras geral do *cluster*/tópico. É interessante ressaltar a variedade de temas ou assuntos abordados: desde segurança da informação; sistemas de saúde; educação, aspectos morais e éticos, redes sociais, sistemas de informação, até saúde pela perspectiva do usuário. As soluções propostas nos artigos estão bastante equilibradas, sendo que metade propõe soluções baseadas em tecnologia e outra metade propõe soluções mais humanas, como elaboração de diretrizes, modelos conceituais e teorias. Entretanto, observa-se que, no aspecto metodológico, a maioria dos artigos utiliza abordagens quantitativas, calcadas na estatística. Um estudo utilizou abordagem qualitativa e um a qualiquantitativa. Ressalta-se que a extração das características dos artigos foi realizada por meio de sua leitura integral.

Quadro 3: Artigos representativos de cada cluster

Tópico	Autor/título	Tema	Abord.	Solução
-1	Não consta (ruído)	Não consta (ruído)		
0	ADJEWA, F.; ESSEGHIR, M.; MERGHEM-	Elaboração de modelo otimizado de LLM para detecção de	Quant.	Tecnologia

	BOULAHIA, L. Semantic Embeddings for Efficient Detection (SEED): a collaborative offloading mechanism for IoT intrusion detection using BERT.	ameaças em ambientes de Internet das coisas (Internet Of Things - IOT).		
1	DAI, Hong-Jie et al. Leveraging large language models for the deidentification and temporal normalization of sensitive health information in electronic health records	Proposta de estratégia de otimização de modelos para desidentificação e normalização de dados clínicos.	Quant.	Tecnologia
2	DANG, Belle et al. Human-AI collaborative learning in mixed reality: Examining the cognitive and socio-emotional interactions.	Identificação de padrões de interação cognitivos e sócio-emocionais na aprendizagem assistida por agente de IA	Quali-quanti.	Modelo/framework conceitual
3	HUI, Bryant P. H. et al. Decoding moral responses in AI: A quantitative analysis of large language models.	Análise de padrões de resposta de modelos de IA a dilemas morais pessoais e impessoais.	Quant.	Proposta de campo de estudo
4	SONG, Inhwa et al. The typing cure: experiences with large language model chatbots for mental health support.	Análise dos padrões de interação e desafios enfrentados por usuários que buscam aconselhamento psicológico com modelos de IA.	Quali.	Modelo/framework conceitual
5	TYNDALL, Erick et al. Feasibility of Local, CPU-Only Large Language Models with Retrieval-Augmented Generation for Secure, Offline Question Answering and Summarization.	Proposta de estratégia para implantação de sistemas de LLM em computadores locais com restrições de recursos e segurança.	Quant.	Tecnologia

6	JIN, Bailu; GUO, Weisi. Synthetic social media influence experimentation via an agentic reinforcement learning large language model bot.	Elaboração de ambiente de simulação com agentes de aprendizado por reforço utilizando LLM para analisar as dinâmicas de influência de opiniões nas redes sociais	Quant.	Tecnologia
7	JARAB, Anan S. et al. Integrating Large Language Models Into UAE Community Pharmacies: Pharmacists' Perspectives on Benefits, Concerns, and Implementation Barriers.	Estudo de percepções e impactos sobre o uso de LLMs de IA na prática profissional de farmacêuticos.	Quant.	Recomendações/ heurísticas

Fonte: Elaborado pelos autores (2026).

A diversidade de assuntos tratados nos artigos reflete, possivelmente, uma parcela da ampla gama de aplicação dos LLMs e seus desafios nos aspectos ético e de privacidade. Notou-se uma ênfase nos temas saúde e educação, o que pode ser explicado pelo seu impacto social, bem como a abundância de dados gerados nessas práticas profissionais, atraindo grandes investimentos monetários (Lopes Júnior, 2025).

O aspecto da Informação em saúde é particularmente relevante levando-se em consideração a necessidade de proteção de dados sensíveis de prontuários de pacientes, sujeitos à regulamentações legais críticas e que podem ter impacto real na vida das pessoas. Analogamente, o emprego da IA no âmbito da educação gera desafios inéditos e complexos para os educadores, que têm buscado novas formas de ensinar, avaliar os discentes, identificar e precaver-se contra o plágio e outras ações antiéticas facilitadas por essa tecnologia. Em vista de tais desafios, criados e/ou ampliados pelas TICs que a ética da informação se faz ainda mais necessária, levando em conta a sua percepção macroscópica da realidade, que tem, como já discutido anteriormente (seção 2), o objetivo último de promover o bem-estar dos agentes informacionais, e porque não dizer, dos sujeitos humanos, e da *infosfera*, ou mesmo da sociedade em si.

É digno de nota que ficaram ausentes nos trabalhos analisados alguns tópicos relevantes, como: direitos autorais, justiça algorítmica, que vem a ser a ausência de vieses discriminatórios nos resultados gerados por IA (Valença; Santos, 2025), e afins. Isto pode indicar uma possível necessidade de maior amadurecimento do tema, que passou a ter um crescimento moderado no número de artigos apenas a partir de 2023, como mostrou a tabela 1.

Os resultados obtidos da análise qualitativa podem ser explicados também pela tímida participação, tanto da Ciência da Informação quanto das Ciências Sociais, no corpus de artigos recuperados. O que parece predominar, portanto, é a visão da ética da computação, com suas análises de contextos e problemas particulares, ao invés da perspectiva ampla da ética da informação. Esta ética da computação se concentra em enxergar a informação por três ângulos principais: como recurso, concernente ao seu uso; como produto, gerando novos recursos informacionais; e no contexto do ambiente informacional, que é afetado por agentes que geram e usam informação (Gonzalez, 2013). Argumenta-se que, embora útil, esta perspectiva traz um entendimento tecnicista e reducionista da realidade informacional.

Argumenta-se que, embora a IA traga problemas prementes que necessitam de regulamentação e discussão urgentes, principalmente nos campos da saúde e da educação, que foram os mais abordados nos artigos, é preciso fomentar discussões éticas sob pressupostos filosóficos mais profundos, amplos e atemporais, como os preconizados por Floridi, Capurro e outros autores. Visto que o universo das TICs está em mutação constante e acelerada, é preciso uma visão ética ‘macro-focada’ que seja válida no longo prazo.

No âmbito da realidade concreta, é interessante notar como princípios universais da ética da informação dão origem a iniciativas de legislações e modelos de governança que regulam o desenvolvimento e uso da IA ao redor do mundo. Um exemplo pioneiro e que vem servindo de modelo para muitos países além de seu âmbito original é o AI Act, desenvolvido pela União Europeia (UE), que trata de aspectos de responsabilização, gestão de risco e preservação da segurança e dignidade humanas relativos ao emprego da IA, seguindo os valores basilares expostos por Capurro (2008). Conforme Kosinski e Scapicchio (2026), essa legislação, que entrou em vigor em 1º de agosto de 2024, classifica os riscos

oferecidos pela IA como de risco inaceitável, alto risco, risco d transparência e risco nulo, atribuindo responsabilidades a toda cadeia de desenvolvimento/uso da IA, desde seus desenvolvedores, empresas terceiras que utilizam suas APIs, e empresas estrangeiras que fornecem produtos/serviços de IA para entidades da UE. Dentre as práticas de IA proibidas, relatam os autores supracitados (2026), estão: realizar classificações discriminatórias de indivíduos com base em seu comportamento social; exploração de pessoas em situação de vulnerabilidade/idosos; sistemas biométricos que identificam sujeitos com base em suas características sensíveis; certas práticas de policiamento preditivo e outras questões afins. Embora o Brasil ainda não tenha uma legislação específica para lidar com a IA, foi elaborado o projeto de lei PL 2338/2023, que se baseia fortemente no IA Act e seu sistema de classificação de risco. Ademais, a Lei Geral de Proteção de Dados (LGPD) brasileira já respalda os direitos individuais quanto ao uso de dados no ambiente da web. Entretanto, D. E. Harris, professor da Escola de Negócios Haas da Universidade da Califórnia em Berkeley, EUA, alerta para possíveis efeitos deletérios da pressão exercida por empresas de tecnologia visando o abrandamento da regulamentação da IA, tanto nos EUA quanto no Brasil, contrapondo o lucro aos direitos dos usuários (Galante, 2026).

A IA, e os LLMs em particular, possivelmente têm produzido tantas lacunas a serem sanadas quanto soluções, e é imprescindível que pesquisadores, estudantes e público em geral, tenham entendimento dessas lacunas éticas e de privacidade, a fim de realizarem escolhas conscientes quanto às ferramentas tecnológicas que utilizam.

5. Considerações finais

Os aspectos de privacidade e ética são fundamentais para o uso seguro dos modelos de LLM. O CVD, por sua vez, se mostra um *frame* adequado para enxergar os aspectos de privacidade dos LLMs no contexto da Ciência da Informação na medida em que a coloca como um elemento ubíquo em todas as fases de acesso aos dados. Este panorama torna-se ainda mais crítico em face de problemas como: vieses

ideológicos; alucinações; vazamento de dados pessoais; conjuntos de dados contendo informações espúrias e assim por diante.

A presente pesquisa revelou uma diversidade de temas e propostas para os problemas éticos e de privacidade em LLMs, mediante a análise das publicações. Em virtude da amplitude de aplicações e da complexidade do tema, vislumbrou-se que soluções tecnológicas podem, possivelmente, ter maior efetividade aliadas ao desenvolvimento de políticas e modelos teóricos que contextualizem e aprofundem a sua aplicação.

No aspecto quantitativo, verificou-se que o tema ética e privacidade em LLMs é emergente e promissor, tendo crescido exponencialmente nos últimos anos. Preocupante, porém é a baixa representação das Ciências Sociais e da Ciência da Informação na produção sobre o tema, o que possivelmente pode refletir uma perspectiva predominantemente tecnicista das discussões. A contribuição de autores e periódicos da América Latina é outro ponto que necessita de atenção urgente para o seu fortalecimento.

Na análise qualitativa, os temas de saúde e educação de destacaram no corpus analisado, tanto pela complexidade dos desafios ora enfrentados no campo informacional nessas áreas, quanto pela sua importância econômica e social. Possivelmente em virtude da participação tímida da Ciência da Informação e das Ciências sociais, é que a abordagem predominante nos trabalhos analisados é da ética micro-focada, com destaque para a ética da computação, tratando de discussões mais contextuais, em detrimento da perspectiva mais ampla da ética da informação. Mostra-se, portanto, imprescindível uma participação mais consistente da Ciência da Informação relativo à privacidade e ética em LLMs, especialmente no âmbito latino-americano, trazendo discussões mais amplas e profundas, que continuem válidas ao longo do constante processo de evolução das TICs.

Referências

BARCAUI, André. (2025). *Guia da inteligência artificial*. São Paulo: Almedina Brasil.

BERTOPIC. (2026). *BERTopic – Advanced Transformer Topic Modeling Library*. Recuperado de https://bertopic.org/

CAPURRO, R. (2006). Privacy. An intercultural perspective. *Ethics and Information* Technology, 8(1), 37–47. Recuperado de https://doi.org/10.1007/s10676-005-4407-4

CAPURRO, R. (2008). Information ethics for and from Africa. *Journal of the American Society for Information Science and Technology*, *59*(7), 1162-1170. Recuperado de https://onlinelibrary.wiley.com/doi/full/10.1002/asi.20850

CAPURRO, R. (2020). Enculturing algorithms. *Ethics and Information Technology*, *22*(2), 131-137. Recuperado de https://link.springer.com/article/10.1007/s11569-019-00340-9?utm_source=researchgate.net&utm_medium=article

CORREIO BRAZILIENSE. (2026, 4 maio). A nova aposta de Mark Zuckerberg: o futuro da inteligência artificial. *Correio Braziliense*. Recuperado de https://www.correiobraziliense.com.br/aqui/2026/05/04/a-nova-aposta-de-mark-zuckerberg-o-futuro-da-inteligencia-artificial/

DUARTE, Danieli Aparecida. (2022). *A ética da informação de Floridi: aplicação aos desafios atuais da Sociedade da Informação para a formação humana* (Dissertação de Mestrado Profissional em Educação Tecnológica). Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro, Campus Uberaba, Uberaba.

FACELI, Katti; LORENA, Ana C.; GAMA, João; ALMEIDA, Tiago Agostinho De; CARVALHO, André C. P. L. F. De. (2025). *Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina* (3. ed.). Rio de Janeiro: LTC.

FEFERBAUM, Marina; SILVA, Alexandre Pacheco da; COELHO, Alexandre Z.; et al. (2023). *Ética, Governança e Inteligência Artificial*. São Paulo: Almedina.

GALANTE, J. P. T. (2026, 16 de junho). *Regulação da IA, interferência das big techs e o futuro pro Brasil*. Instituto de Estudos Avançados da Universidade de São Paulo. Recuperado de https://www.iea.usp.br/noticias/regulacao-da-ia-interferencia-das-big-techs-e-o-futuro-pro-brasil

GONZALEZ, Maria Nélida. (2013, jan./jun.). Luciano Floridi e os problemas filosóficos da informação: da representação à modelização. *InCID*, *4*(1), 03-25. Recuperado de https://cidad.bu.ufsc.br/files/2021/10/%C3%89tica-de-Floridi-por-N%C3%A9lida-Gonzales.pdf

IBM. (2025). *Risco de ataque de inferência de associação para IA*. Recuperado de https://www.ibm.com/docs/pt-br/watsonx/saas?topic=atlas-membership-inference-attack

KOSINSKI, M., & SCAPICCHIO, M. (n.d.). *What is the EU AI Act?* IBM. Recuperado em maio 26, 2026, de https://www.ibm.com/think/topics/eu-ai-act

KRANTZ, Tom; JONKER, Alexandra. (2025). *Jailbreak de IA: desenraizar uma ameaça em evolução*. IBM Think Insights. Recuperado de https://www.ibm.com/br-pt/think/insights/ai-jailbreak

LABATE, Alice. (2025, 27 maio). Meta anuncia divisão de AGI, 'super' inteligência artificial, de olho em rivais como OpenAI e Google. *O Estado de São Paulo*. Recuperado de https://www.estadao.com.br/link/empresas/meta-anuncia-divisao-de-agi-super-inteligencia-artificial-de-olho-em-rivais-como-openai-e-google-nprei/

LOPES JUNIOR, José Evaldo Gonçalves et al. (2025). Tecnologias digitais e inteligência artificial na educação em saúde: transformações do ambiente acadêmico ao contexto hospitalar. *Revista Educação & Ensino, 9*. Recuperado de https://periodicos.uniateneu.edu.br/index.php/revista-educacao-e-ensino/article/view/882/534

MARTINELLI, Dante P. (2012). *TEORIA GERAL DOS SISTEMAS*. Rio de Janeiro: Saraiva.

NIST (National Institute of Standards and Technology). (2024, jul.). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1). Gaithersburg, MD: National Institute of Standards and Technology. Recuperado de https://doi.org/10.6028/NIST.AI.600-1

OCDE. (2025). *How can AI help make better regional development policies in Korea* (OECD Regional Development Papers). Paris: OECD Publishing. Recuperado de https://www.oecd.org/en/publications/how-can-ai-help-make-better-regional-development-policies-in-korea_be01f52d-en.html

OLFF, Sabine. (2025, 1 dez.). *As alternativas europeias ao ChatGPT* (vídeo). Brasil: Deutsch Welle [DW]. Recuperado de https://www.dw.com/pt-br/as-alternativas-europeias-ao-chatgpt/video-74962113

PELLEGRINI, Eliane; VITORINO, Elizete Vieira. (2018, abr./jun.). A dimensão ética da competência em informação sob a perspectiva da filosofia. *Perspectivas em Ciência da Informação, 23*(2), 117-133. doi: https://doi.org/10.1590/1981-5344/2953

RUSSELL, Stuart J.; NORVIG, Peter. (2022). *Inteligência Artificial: Uma Abordagem Moderna* (4. ed.). Rio de Janeiro: GEN LTC.

RUSSELL, Stuart J.; NORVIG, Peter. (2022). *Inteligência Artificial - Uma Abordagem Moderna* (4. ed., e-book). Rio de Janeiro: GEN LTC.

SANT'ANA, Ricardo César Gonçalves. (2016, maio/ago.). Ciclo de vida dos dados: uma perspectiva a partir da ciência da informação. *Informação & Informação, 21*(2), 116-142. doi: 10.5433/1981-8920.2016v21n2p116

SANT'ANA, Ricardo César Gonçalves; MARTINS, Dayane de Oliveira. (2024, set./dez.). Contextualizando a taxonomia da privacidade de Solove no ciclo de vida dos dados. *Comun. Mídia Consumo*, *21*(62), 558-582. doi: 10.18568/cmc.v20162.2873

SCHAU, Kevin; CHEN, Szu Yu; TIKU, Nitasha. (2023, 19 abr.). Inside the secret list of websites that make AI like ChatGPT sound smart. *The Washington Post*. Recuperado de https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

SHANAHAN, Murray; MCDONELL, Kyle; REYNOLDS, Laria. (2023, 16 nov.). Role play with large language models. *Nature*, *623*, 493–498. doi: https://doi.org/10.1038/s41586-023-06647-8

VALENÇA, Lucas Rodrigues; SANTOS, Ronnie de Souza. (2025). Justiça algorítmica: instrumentalização, limites conceituais e desafios na engenharia de software. In: *Anais do 6. Workshop sobre as Implicações da Computação na Sociedade (WICS)*, Maceió, p. 225-234. Porto Alegre: Sociedade Brasileira de Computação. doi: https://doi.org/10.5753/wics.2025.8032

VASWANI, A. et al. (2017). Attention Is All You Need. In: *Anais da 31. Conference on Neural Information Processing Systems (NIPS)*, Long Beach. Long Beach: NIPS. Recuperado de https://arxiv.org/abs/1706.03762

YE, Jiayuan et al. (2022). Enhanced membership inference attacks against machine learning models. In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*, Los Angeles, CA, USA, p. 1-14. New York, NY, USA: ACM. doi: https://doi.org/10.1145/3548606.3560675

ZHAO, Pinlong et al. (2025). *Data Poisoning in Deep Learning: A Survey*. Recuperado de https://github.com/Pinlong-Zhao/Data-Poisoning

ZOU, Andy; WANG, Zifan; KOLTER, J. Zico; FREDRIKSON, Matt. (2023). *Universal and transferable adversarial attacks on aligned language models*. Recuperado de https://arxiv.org/abs/2307.15043

Nota del editor

El editor responsable por la publicación de este trabajo es Yanet Fuster

Nota de disponibilidad de datos

El conjunto de datos que apoya los resultados de este estudio se encuentran disponibles em https://github.com/lagostinellipps/posdoc.

Nota dos autores

O presente artigo foi desenvolvido no âmbito do programa de pós-doutorado da Universidade estadual Paulista – UNESP, Brasil.