26º ENANCIB
Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
A Amazônia na sociedade global em rede: assimetrias e potencialidades informacionais, tecnológicas e ambientais
Belém, Pará, 09 a 13 de novembro de 2026
Universidade Federal do Pará

ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
ISSN 2177-3688

GT 8 – Dados, Informação e Tecnologia

***LARGE LANGUAGE MODELS* EM REVISÕES SISTEMÁTICAS E SÍNTESE DE EVIDÊNCIAS:** UM MAPEAMENTO BIBLIOMÉTRICO TEMAC DAS APLICAÇÕES, DO DESEMPENHO E DOS NÍVEIS DE AUTOMAÇÃO

***LARGE LANGUAGE MODELS IN SYSTEMATIC REVIEWS AND EVIDENCE SYNTHESIS**: A TEMAC BIBLIOMETRIC MAPPING OF APPLICATIONS, PERFORMANCE AND AUTOMATION LEVELS*

Marília Augusta de Freitas – Universidade de Brasília (UnB)
Ari Melo Mariano – Universidade de Brasília (UnB)

Modalidade: Trabalho Completo

Resumo: A adoção de *Large Language Models* (LLMs) em revisões sistemáticas e sínteses de evidências cresce rapidamente, mas as evidências empíricas permanecem dispersas e persistem dúvidas sobre confiabilidade, reprodutibilidade e desempenho. Este estudo sistematizou e analisou essas evidências, identificando aplicações, desempenho, benefícios, limitações e níveis de automação. Adotou-se uma abordagem mista, integrando a Teoria do Enfoque Metanalítico Consolidado (TEMAC) ao protocolo PRISMA. Dos 744 registros recuperados na Scopus e na Web of Science (2023–2026), 106 compuseram o corpus bibliométrico e 95 textos completos fundamentaram a síntese qualitativa. As aplicações concentram-se na triagem e na extração de dados; a família GPT predomina; e a avaliação utiliza principalmente *recall*, precisão, F1 e concordância humana. Os resultados indicam alto *recall*, precisão mais variável e necessidade de supervisão humana. Como contribuição, propõe-se um modelo integrador estruturado em tarefas, tecnologia, avaliação e governança.
Palavras-chave: *Large Language Models*; revisão sistemática; síntese de evidências; TEMAC; bibliometria.

Abstract: The adoption of Large Language Models (LLMs) in systematic reviews and evidence synthesis is growing rapidly, but empirical evidence remains fragmented and concerns persist regarding reliability, reproducibility and performance. This study systematized and analysed this evidence, identifying applications, performance, benefits, limitations and automation levels. A mixed approach integrated the Consolidated Meta-Analytical Approach Theory

(TEMAC) with the PRISMA protocol. Of 744 records retrieved from Scopus and Web of Science (2023–2026), 106 composed the bibliometric corpus and 95 full-text studies supported the qualitative synthesis. Applications concentrate on screening and data extraction; the GPT family predominates; and evaluation relies mainly on recall, precision, F1 and human agreement. Results indicate high recall, more variable precision and continued need for human supervision. The main contribution is an integrative model structured around tasks, technology, evaluation and governance.
Keywords: Large Language Models; systematic review; evidence synthesis; TEMAC; bibliometrics.

1 INTRODUÇÃO

A síntese de evidências constitui um dos pilares da pesquisa contemporânea, mas o crescimento acelerado da literatura científica tem ampliado o esforço necessário para identificar, selecionar e sintetizar estudos relevantes. Nesse contexto, *Large Language Models (LLMs)*, como GPT, *Claude, Gemini, Llama* e *Mistral*, passaram a ser utilizados para apoiar ou automatizar etapas das revisões sistemáticas, especialmente triagem de referências e extração de dados.

Embora estudos recentes apontem ganhos de produtividade e redução do esforço humano, persistem dúvidas sobre confiabilidade, reprodutibilidade e desempenho desses modelos quando comparados a revisores humanos ou a métodos tradicionais de automação. Além disso, as evidências encontram-se dispersas entre diferentes áreas, dificultando uma visão consolidada sobre aplicações, desempenho, benefícios, limitações e níveis de automação alcançados.

Diante dessa lacuna, este estudo tem como objetivo sistematizar as evidências empíricas disponíveis na literatura e analisar suas inter-relações sobre o uso de LLMs em revisões sistemáticas e sínteses de evidências. A questão principal investiga quais evidências empíricas existem sobre o uso de LLMs para apoiar ou automatizar etapas das revisões sistemáticas, desdobrando-se em sete perguntas relativas às etapas apoiadas, aos modelos utilizados, às métricas de avaliação, ao desempenho em relação aos revisores humanos, aos benefícios reportados, às limitações identificadas e aos níveis de automação alcançados.

Para responder a essas questões, adotou-se uma abordagem mista, integrando a Teoria do Enfoque Metanalítico Consolidado (TEMAC) ao protocolo PRISMA. A análise bibliométrica permitiu caracterizar a evolução do campo, sua base intelectual e suas frentes de pesquisa, enquanto a síntese qualitativa dos textos completos respondeu às perguntas relativas às

aplicações, aos modelos, às métricas, ao desempenho, aos benefícios, às limitações e aos níveis de automação.

Os resultados evidenciam concentração das aplicações na triagem de referências e na extração de dados, desempenho variável conforme a tarefa e predomínio de fluxos conduzidos com supervisão humana, achados que sustentam a proposição de um modelo integrador estruturado nas dimensões de tarefas, tecnologia, avaliação e governança, discutido à luz dos referenciais teóricos sobre automação e trabalho intelectual apresentados na seção seguinte.

2 REFERENCIAL TEÓRICO

As revisões sistemáticas caracterizam-se pelo uso de protocolos explícitos, critérios de elegibilidade predefinidos e procedimentos documentados de busca, seleção, avaliação e síntese dos estudos, garantindo transparência e replicabilidade (Higgins *et al.*, 2023; Page *et al.*, 2021). Embora esse rigor metodológico seja fundamental para a confiabilidade das evidências, ele também impõe elevada carga operacional aos revisores. Nesse contexto, a automação passou a ser empregada inicialmente em tarefas específicas, especialmente na triagem de estudos por técnicas de aprendizado de máquina. O surgimento de uma nova geração de modelo ampliou esse cenário ao possibilitar a aplicação de uma mesma tecnologia em múltiplas etapas do fluxo de revisão.

Entretanto, a literatura relata, de forma pontual, limitações para seu uso, incluindo alucinações, sensibilidade aos prompts, dificuldades de reprodutibilidade e menor precisão em tarefas que exigem raciocínio quantitativo complexo (Mahmoudi *et al.*, 2025; Motzfeldt Jensen *et al.*, 2025; Nordmann; Fischer, 2025; Oami; Okada; Maeda; Nakada, 2026; Yisha *et al.*, 2026). Tais restrições, contudo, permanecem dispersas entre estudos que examinam modelos, etapas ou domínios isolados, o que torna indispensável consolidar e analisar criticamente essas evidências antes de admitir a incorporação rotineira dos LLMs aos processos de síntese de evidências. Essa consolidação demonstra que tal incorporação envolve não apenas questões de desempenho, mas também decisões relacionadas ao controle, à supervisão e à responsabilização pelos resultados produzidos.

Além das questões técnicas, a adoção de LLMs em revisões sistemáticas suscita discussões sobre a redistribuição de atividades entre pesquisadores e sistemas automatizados. Estudos recentes indicam que tarefas como avaliação crítica da evidência, interpretação dos resultados e formulação de conclusões continuam dependentes do julgamento humano,

mantendo os pesquisadores como responsáveis pela qualidade e validação dos resultados (Motzfeldt Jensen *et al.*, 2025; Sollini *et al.*, 2025). Essa discussão dialoga com debates mais amplos sobre automação e trabalho intelectual. Enquanto autores como Braverman (1974) e Zuboff (2019) enfatizam os riscos associados à transferência de competências humanas para sistemas técnicos, Davenport e Kirby (2016) propõem o conceito de aumentação (*augmentation*), segundo o qual a tecnologia complementa o trabalho humano ao assumir tarefas repetitivas e de larga escala, preservando aos profissionais as atividades que exigem interpretação, julgamento e responsabilidade. Essa perspectiva orienta o presente estudo, que busca compreender em que medida as evidências empíricas sobre o uso de LLMs em revisões sistemáticas apontam para processos de substituição, complementaridade ou aumentação do trabalho dos revisores, bem como os mecanismos de governança que condicionam esses diferentes níveis de automação.

No campo da Ciência da Informação, essa problemática envolve a qualidade, a confiabilidade, a organização, a mediação e a governança da informação científica. Ao modificar os processos de seleção, representação e validação das evidências, o uso de LLMs situa-se diretamente no escopo dos estudos sobre dados, informação e tecnologia.

3 PROCEDIMENTOS METODOLÓGICOS

Para sistematizar as evidências dispersas na literatura e analisar suas inter-relações, o estudo caracteriza-se como uma revisão de literatura de abordagem mista, na qual a dimensão bibliométrica e a síntese qualitativa cumprem funções complementares: a primeira sistematiza a evolução, a estrutura intelectual e as frentes do campo, enquanto a segunda analisa as inter-relações entre aplicações, modelos, desempenho, benefícios, limitações e níveis de automação, respondendo diretamente às sete perguntas de pesquisa a partir dos textos completos disponíveis. A dimensão bibliométrica foi conduzida pela TEMAC, estruturada nas etapas de preparação da busca, apresentação e inter-relação dos dados e detalhamento por cocitação e acoplamento (Mariano; Rocha, 2017). A seleção e o relato seguiram o fluxo PRISMA 2020 (Page *et al*., 2021), e o mapeamento dos campos intelectual e de fronteira utilizou o *VOSviewer* (Van Eck; Waltman, 2010). Desse modo a pesquisa se caracteriza por ser exploratória com abordagem qualitativa.

As buscas foram realizadas na Scopus e na Web of Science Core Collection, bases de cobertura ampla e curadoria reconhecida para estudos bibliométricos. A estratégia combinou

quatro blocos conceituais conectados por operadores booleanos: tipo de síntese de evidências, etapas e ferramentas de revisão, técnicas de inteligência artificial e termos de avaliação de desempenho, com cláusula de exclusão para estudos puramente bibliométricos, revisões guarda-chuva ou estudos de predição de risco. O recorte de 2023 a 2026 foi definido em razão da difusão pública dos LLMs generativos a partir do final de 2022 e do crescimento de suas aplicações nas revisões sistemáticas. Aplicaram-se filtros de tipo documental para artigos e trabalhos de eventos. O Quadro 1 apresenta a estratégia sintética utilizada em cada base.

Quadro 1 - Estratégia de busca e registros recuperados por base de dados

Base de dados	Estratégia de busca	Filtros	Registros
Scopus	*TITLE-ABS-KEY: (evidence synthesis / systematic, scoping review) AND (screening, selection, data extraction, review tools) AND (machine learning, NLP, LLM, GPT, Claude, AI) AND (performance, accuracy, recall, precision, F1, agreement); NOT bibliometric/umbrella/risk prediction*	2023-2026; artigos e trabalhos de eventos	463
Web of Science	TS = mesmos quatro blocos conceituais e mesma cláusula de exclusão (NOT) utilizados na estratégia anterior	2023-2026; artigos e trabalhos de eventos	281
Total recuperado	Corpus integrado das duas bases de dados	—	744

Fonte: elaboração própria (2026).

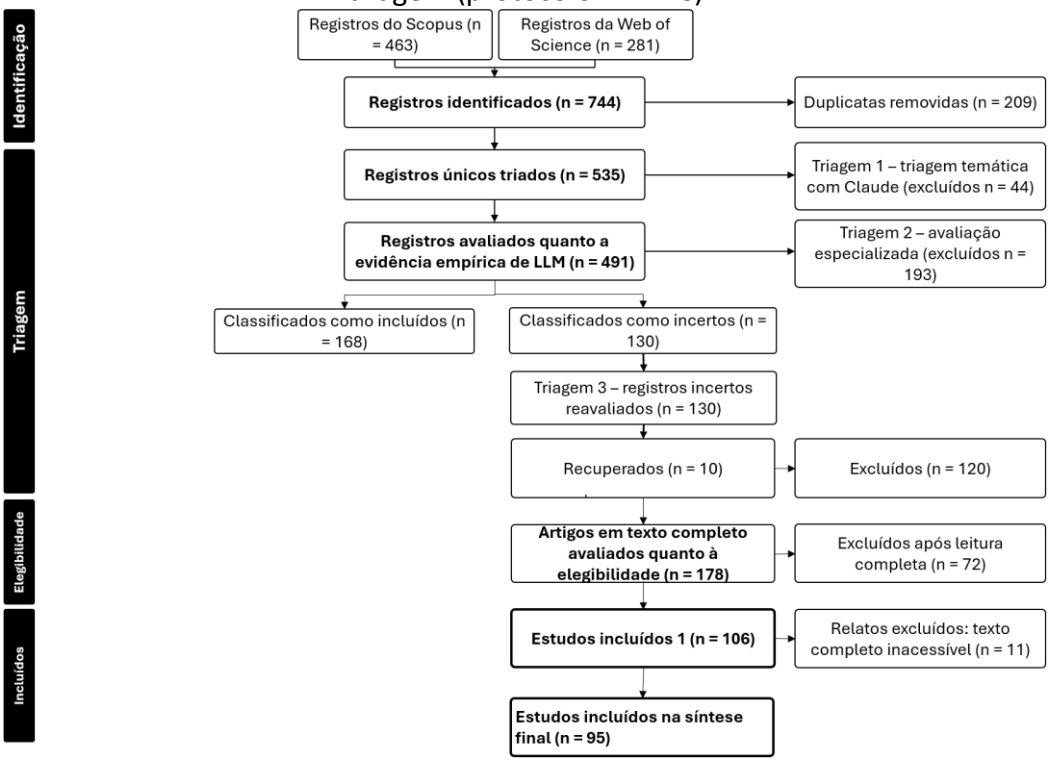
A seleção dos estudos compreendeu três etapas sucessivas de triagem, seguidas pela avaliação de elegibilidade em texto completo (Figura 1), todas registradas em planilhas auditáveis. Após a remoção de duplicatas, o modelo Claude foi empregado exclusivamente como apoio à classificação preliminar. Os registros classificados como potencialmente elegíveis ou incertos foram submetidos a avaliações especializadas e validação humana, permanecendo sob responsabilidade dos pesquisadores a decisão final de inclusão no corpus. O processo resultou em 106 estudos incluídos. A concordância entre os avaliadores humanos foi monitorada por meio do coeficiente Kappa, conforme Landis e Koch (1977).

Foram incluídos estudos que avaliassem empiricamente o uso de LLMs em etapas das revisões sistemáticas, apresentando métricas quantitativas de desempenho ou procedimentos explícitos de validação. Estudos sobre aprendizado de máquina clássico e ferramentas dedicadas foram mantidos apenas quando utilizados como linhas de base comparativas para a

avaliação dos LLMs. Foram excluídos estudos sem avaliação empírica, aplicações não relacionadas à síntese de evidências e documentos sem resultados originais.

Dos 106 estudos incluídos no corpus bibliométrico, 95 possuíam texto completo disponível e constituíram a base da síntese qualitativa. Para cada estudo, foram registradas informações sobre etapas apoiadas, modelos e ferramentas, métricas de avaliação, desempenho em relação a revisores humanos, benefícios, limitações, níveis de automação e características dos prompts, quando disponíveis.

Figura 1 - Fluxograma PRISMA do processo de seleção dos estudos em três etapas de triagem (protocolo TEMAC)



Fonte: elaboração própria (2026).

As categorias foram definidas a partir das perguntas de pesquisa e utilizadas na construção do modelo integrador. A síntese por família de modelo, área, etapa, tipo de estudo, benefícios, limitações e divulgação dos prompts encontra-se detalhada nos quadros 4 e 5 do Apêndice A.

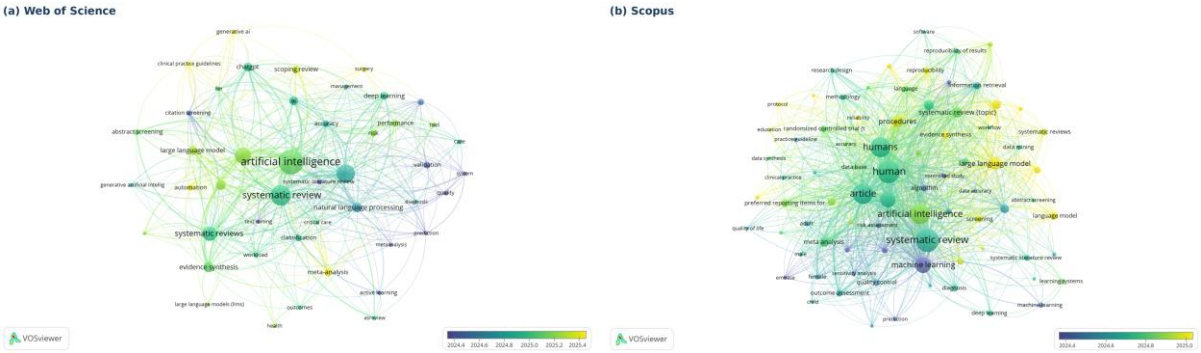
4 RESULTADOS: PANORAMA BIBLIOMÉTRICO E DELIMITAÇÃO DO CORPUS

A produção sobre LLMs em revisões sistemáticas é recente e fortemente ascendente. Na Web of Science, os registros passaram de 15 em 2023 para 127 em 2025, enquanto na *Scopus* evoluíram de 57 para 198 no mesmo período. O campo concentra-se em periódicos de metodologia de pesquisa e informática biomédica, com destaque para *PLoS ONE*, *BMJ Open*,

Systematic Reviews, *Research Synthesis Methods* e *Journal of Medical Internet Research*. A produção é liderada por Estados Unidos, Canadá, Reino Unido e China, com participação crescente de Austrália, Alemanha, Países Baixos, Japão e Brasil. Além do crescimento quantitativo, os mapas de coocorrência permitem compreender como os temas e conceitos do campo vêm se reorganizando ao longo do tempo.

A leitura dos mapas de coocorrência de palavras-chave (Figura 2) confirma essa transformação. Os termos de maior densidade, inteligência artificial, revisão sistemática e síntese de evidências, articulam-se a um núcleo metodológico mais antigo, formado por mineração de texto, processamento de linguagem natural, aprendizado ativo e ferramentas como *ASReview*, e a um núcleo recente, organizado em torno de *large language model*, *ChatGPT*, IA generativa, extração de dados e *workflow*. A coloração temporal evidencia que os termos generativos emergem sobretudo a partir de 2025, deslocando o vocabulário do campo da automação estatística para a automação cognitiva.

Figura 2 - Mapas de coocorrência de palavras-chave com sobreposição temporal. (a) *Web of Science*; (b) *Scopus*



Fonte: elaboração própria (2026).

A codificação dos 95 estudos com texto completo respondeu diretamente às perguntas secundárias. Quanto às etapas apoiadas (PQ1), as aplicações concentraram-se na extração de dados e na triagem de títulos e resumos, seguidas por fluxos completos, avaliação de risco de viés e triagem de texto completo. Quanto aos modelos (PQ2), a família GPT esteve presente em 80 estudos, seguida por Claude (n=22), Gemini (n=18), Llama (n=14) e Mistral (n=7); um mesmo estudo poderia avaliar mais de uma família. Abordagens de aprendizado de máquina clássico e ferramentas como *ASReview*, *Rayyan* e *RobotReviewer* foram utilizadas como linhas de base comparativas. As métricas (PQ3) privilegiaram revocação ou sensibilidade, precisão, F1, economia de trabalho (WSS), acurácia e concordância com humanos.

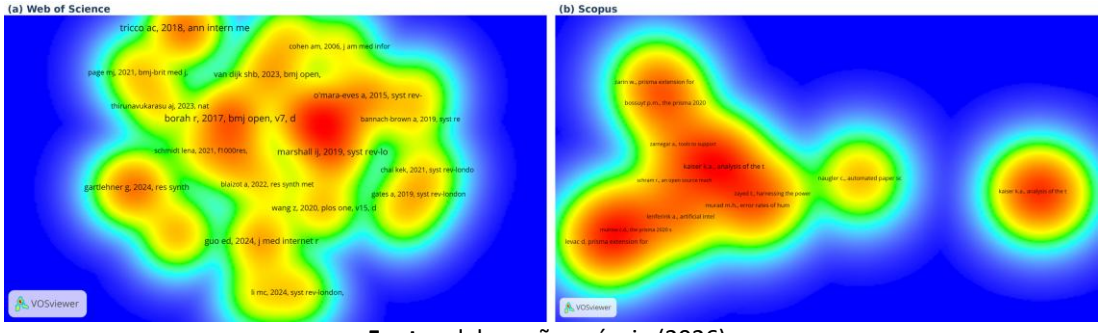
As análises bibliométricas consideraram os 106 estudos incluídos, enquanto a síntese sobre etapas, modelos, métricas, desempenho, benefícios, limitações e níveis de automação baseou-se nos 95 textos completos. Essa distinção evita interpretar os mapas bibliométricos como substitutos da análise das evidências empíricas. O detalhamento complementar da codificação está disponível no Apêndice A.

4.1 BASE INTELECTUAL: ANÁLISE DE COCITAÇÃO

A análise de cocitação revela os fundamentos sobre os quais o campo se assenta. Por opção metodológica, a leitura foi qualitativa a partir dos nós dos mapas (Figura 3), identificando três escolas convergentes nas duas bases. A primeira reúne os padrões de relato e condução de revisões sistemáticas, representados por Page *et al.* (2021), Tricco *et al.* (2018) e Borah *et al.* (2017), responsáveis por estabelecer normas metodológicas, transparência e enquadramento da carga de trabalho das revisões. A segunda corresponde à mineração de texto e ao aprendizado de máquina para triagem, ancorada em O'Mara-Eves *et al.* (2015), Marshall e Wallace (2019) e Bannach-Brown *et al.* (2019), que forneceram métricas de avaliação, desenhos de validação e fundamentos para a automação das revisões. A terceira representa a virada generativa dos LLMs, associada a van Dijk *et al.* (2023), Thirunavukarasu *et al.* (2023) e aos primeiros estudos de prova de conceito, que reconfiguraram atividades como triagem e extração de dados em tarefas baseadas em linguagem natural. A interpretação considerou a centralidade dos nós, a intensidade dos vínculos, a formação dos agrupamentos e a leitura dos documentos mais representativos de cada *cluster*, reduzindo a dependência de inspeção exclusivamente visual dos mapas.

No mapa da Scopus, o mesmo arranjo aparece com a presença adicional de marcos de relato ao lado de revisões sobre ferramentas de automação, reforçando a dupla ancoragem do campo: metodológica e tecnológica.

Figura 3 - Mapas de densidade de cocitação da base intelectual. (a) *Web of Science*; (b) *Scopus*

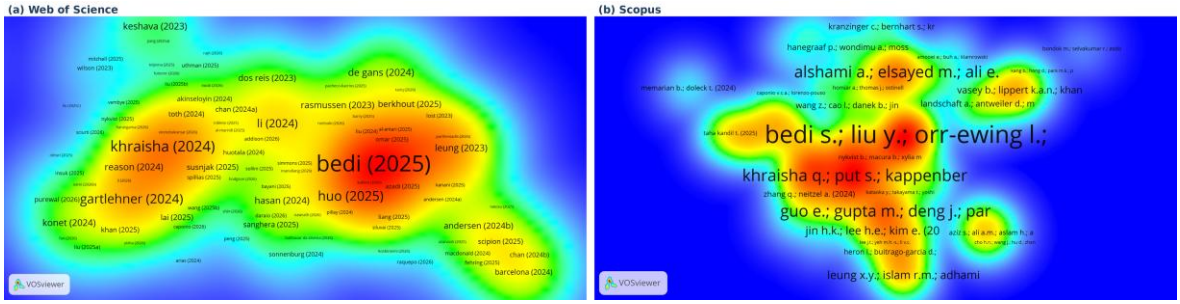


Fonte: elaboração própria (2026).

4.2 FRENTES DE PESQUISA: ACOPLAMENTO BIBLIOGRÁFICO

O acoplamento bibliográfico identifica as frentes ativas, isto é, os documentos recentes que compartilham referências e, portanto, problemas comuns. Os mapas (Figura 4) revelam forte convergência entre as bases: os documentos de maior conectividade são, em ambas, estudos de 2024 e 2025 que avaliam empiricamente LLMs em tarefas específicas. Destacam-se Khraisha *et al.* (2024), que testa o GPT em múltiplas etapas; Gartlehner *et al.* (2024) na extração de dados; Guo *et al.* (2024) e Li, Sun e Tan (2024) na triagem de resumos; e Sanghera *et al.* (2025), com ensembles de modelos para triagem de alto desempenho.

Figura 4 - Mapas de densidade de acoplamento bibliográfico das frentes de pesquisa. (a) *Web of Science*; (b) *Scopus*



Fonte: elaboração própria (2026).

Organizando os documentos por afinidade temática, distinguem-se três eixos. O eixo de seleção e triagem agrupa Akinseloyin*,* Jiang e Palade (2024), Oami, Okada e Nakada (2025) e Sanghera *et al.* (2025), voltados à automação de título e resumo com foco em revocação. O eixo de extração e síntese reúne Gartlehner *et al.* (2024), Spillias *et al.* (2025), Reason *et al.* (2024) e Susnjak *et al.* (2025), que avançam da extração de PICOs ao ajuste fino de modelos de domínio. O eixo de *workflow* integral, com Hasan *et al.* (2024) e Alshami *et al.* (2023), propõe arquiteturas que encadeiam várias etapas sob supervisão humana. O quadro 2 resume as frentes.

Quadro 2 - Frentes de pesquisa identificadas por meio do acoplamento bibliográfico

Frente de pesquisa	Estudos representativos
Seleção e triagem	Akinseloyin*,* Jiang e Palade (2024); Oami, Okada e Nakada (2025); Sanghera *et al.* (2025)
Extração e síntese de dados	Gartlehner *et al.* (2024); Spillias *et al.* (2025); Reason *et al.* (2024); Susnjak *et al.* (2025)
Fluxo de trabalho ponta a ponta	Hasan *et al.* (2024); Alshami *et al.* (2023); Khraisha *et al.* (2024)

Fonte: elaboração própria (2026).

A síntese entre base e frentes evidencia uma trajetória clara: as frentes recentes herdam das escolas clássicas as métricas de avaliação e os procedimentos de validação, mas reposicionam o problema da automação estatística para a automação generativa, na qual o

mesmo modelo executa, por instrução em linguagem natural, tarefas antes resolvidas por algoritmos especializados.

4.3 SÍNTESE DAS EVIDÊNCIAS EMPÍRICAS E MODELO INTEGRADOR

A articulação entre a síntese dos 95 textos completos, o panorama bibliométrico, a base intelectual e as frentes de pesquisa permitiram propor um modelo integrador do campo, organizado em quatro dimensões interdependentes (Figura 5). A dimensão de tarefas descreve as etapas apoiadas; a tecnológica, os modelos e arquiteturas mobilizados; a de avaliação, as métricas, os desenhos de validação e os padrões de desempenho; e a de governança, os benefícios, os riscos e os graus de autonomia. O construto central, síntese de evidências assistida por LLM, emerge da interação entre essas dimensões.

Figura 5 - Modelo integrador do uso de LLMs em revisões sistemáticas e síntese de evidências



Fonte: elaboração própria (2026).

O modelo integrador sintetiza os achados empíricos e bibliométricos ao evidenciar que a incorporação de LLMs às revisões sistemáticas não depende apenas da disponibilidade de modelos mais avançados, mas da articulação entre tarefas, tecnologias, evidências de desempenho e mecanismos de governança. A expansão das aplicações impulsiona o desenvolvimento de novos modelos; os avanços tecnológicos demandam estratégias de avaliação mais robustas; e os resultados dessas avaliações condicionam os níveis aceitáveis de autonomia. As respostas agregadas às perguntas de pesquisa são apresentadas no Quadro 3, e o detalhamento por família de modelo encontra-se no Apêndice A.

Quadro 3 - Síntese das respostas às perguntas de pesquisa

PQ	Foco	Resposta baseada em evidências
PQ1	Etapas da revisão apoiadas	Predominantemente extração de dados e triagem de título/resumo; depois fluxos completos, risco de viés e texto completo
PQ2	Modelos e ferramentas	Família GPT dominante; depois Claude, Gemini, Llama, Mistral; ML clássico e ferramentas dedicadas como linha de base
PQ3	Métricas de avaliação	*Recall*/sensibilidade, precisão, F1, WSS, acurácia e concordância humana (*kappa*)
PQ4	Desempenho vs. Humanos	Na triagem, alto *recall* e precisão mais variável; adequado como primeiro filtro, com validação humana
PQ5	Benefícios relatados	Grande redução de carga de trabalho e tempo, acessibilidade, explicabilidade das decisões
PQ6	Limitações e riscos	Alucinação, critérios de inclusão complexos, dependência dos dados de treinamento, baixa reprodutibilidade, viés
PQ7	Nível de automação	Predominantemente apoio com humano no circuito (*human-in-the-loop*); automação plena apenas em tarefas restritas e com mecanismos de auditoria

Fonte: elaboração própria (2026).

A síntese confirma a interação entre as quatro dimensões do modelo integrador. Na dimensão de tarefas, a automação permanece concentrada na triagem de estudos e na extração de dados, mas avança progressivamente para fluxos de trabalho mais amplos, evolução já refletida nas três frentes de pesquisa descritas no Quadro 2. Na dimensão tecnológica, a predominância da família *GPT*, seguida por *Claude*, *Gemini*, *Llama* e *Mistral*, confirma que os LLMs se tornaram a principal infraestrutura tecnológica do campo, resultado da convergência entre os padrões metodológicos das revisões sistemáticas, as abordagens de mineração de texto e aprendizado de máquina e a virada generativa representada pelos primeiros estudos de prova de conceito, discutidos na base intelectual apresentada na seção 4.2.

A dimensão da avaliação revelou que o desempenho varia conforme a tarefa. Na triagem de títulos e resumos, os estudos indicaram predominantemente alto *recall*, acompanhado de precisão mais variável, o que favorece o uso dos modelos como primeiro filtro, mas mantém a necessidade de revisão humana. Na extração de dados, o desempenho foi mais consistente para

campos estruturados e mais sensível à complexidade das variáveis, à extensão dos documentos, à formulação dos prompts e à necessidade de raciocínio quantitativo. Esses resultados explicam tanto os benefícios reportados, redução da carga de trabalho, economia de tempo e maior acessibilidade, quanto as limitações recorrentes, como alucinação, viés e baixa reprodutibilidade. A governança, portanto, não se reduz à presença de um humano no circuito: requer registro dos prompts, identificação das versões dos modelos, documentação dos parâmetros, preservação das entradas e saídas, auditoria e definição da responsabilidade pelas decisões. O principal resultado não é a demonstração de viabilidade irrestrita, mas a constatação de que o uso confiável de LLMs depende do equilíbrio entre desempenho, transparência, supervisão e responsabilização.

4.4 CONTRIBUIÇÕES PARA A CIÊNCIA DA INFORMAÇÃO

Como contribuição teórica, o estudo evidencia a convergência entre a tradição da mineração de texto e do aprendizado de máquina e a recente virada generativa dos LLMs; como contribuição prática, oferece um referencial para avaliar oportunidades e limites da automação nas diferentes etapas da síntese de evidências. Para a Ciência da Informação, os achados trazem ganhos em quatro frentes complementares: na organização e representação da informação, os LLMs passam a atuar sobre operações clássicas do campo - seleção, indexação, condensação e representação de conteúdo -, antes dependentes do julgamento individual do revisor; na recuperação e avaliação da informação, o padrão de alto recall com precisão mais variável recoloca, em nova base tecnológica, a tensão entre exaustividade e pertinência, mantendo indispensáveis as métricas já consolidadas na área; na mediação, o predomínio de fluxos com humano no circuito reafirma o papel do profissional da informação como mediador qualificado, responsável por formular prompts, delimitar critérios e validar decisões; e, na governança e na ética informacional, limitações recorrentes como alucinações, vieses e baixa reprodutibilidade convertem-se em exigências de registro de prompts, identificação de versões e atribuição de responsabilidade, requisitos de curadoria e procedência documental já consolidados na área. O modelo integrador proposto oferece, assim, um vocabulário comum - tarefas, tecnologia, avaliação e governança - para investigar, ensinar e regular o uso de LLMs na síntese de evidências.

5 CONSIDERAÇÕES FINAIS

O problema que orientou esta pesquisa foi assim formulado: “quais evidências empíricas existem sobre o uso de LLMs para apoiar ou automatizar etapas das revisões sistemáticas”, desdobrado em sete perguntas relativas às etapas apoiadas, aos modelos utilizados, às métricas de avaliação, ao desempenho em relação aos revisores humanos, aos benefícios reportados, às limitações identificadas e aos níveis de automação alcançados. A resposta é que já existe um corpo de evidências consistente, embora assimétrico: predominam a triagem de título/resumo e a extração de dados, apoiadas sobretudo na família GPT, com padrão de alto recall e precisão mais variável - o que sustenta o uso dos modelos como primeiro filtro e segundo revisor, não como substitutos do revisor humano. Os benefícios relatados convivem com limitações recorrentes como alucinação, viés e baixa reprodutibilidade, mantendo os níveis de automação predominantemente de apoio com humano no circuito.

O objetivo geral, assim enunciado, “sistematizar as evidências empíricas disponíveis na literatura e analisar suas inter-relações sobre o uso de LLMs em revisões sistemáticas e sínteses de evidências”, foi alcançado por meio da TEMAC articulada ao PRISMA. A cocitação e o acoplamento bibliográfico explicitaram a base intelectual e as três frentes do campo - seleção e triagem, extração e síntese de dados, e fluxos de trabalho ponta a ponta -, enquanto a síntese qualitativa articulou aplicações, modelos, métricas, desempenho, benefícios, limitações e níveis de automação no modelo integrador estruturado nas dimensões tarefas, tecnologia, avaliação e governança.

Entre as limitações, destacam-se o recorte temporal recente e a interpretação qualitativa dos mapas, mitigada pela consideração conjunta da estrutura das redes e do conteúdo dos estudos representativos. Pesquisas futuras podem aprofundar o desempenho por tarefa, realizar meta-análises quantitativas e desenvolver mecanismos de auditoria, rastreabilidade, governança e responsabilização algorítmica.

REFERÊNCIAS

AKINSELOYIN, Olalekan; JIANG, Xiaorui; PALADE, Vasile. A question-answering framework for automated abstract screening using large language models. **Journal of the American Medical Informatics Association**, [*s. l.*], v. 31, n. 9, p. 1939-1952, 2024. Disponível em: https://doi.org/10.1093/jamia/ocae166. Acesso em: 6 nov. 2026.

ALSHAMI, Ahmad; ELSAYED, Moustafa; ALI, Eslam; ELTOUKHY, Abddelrahman *et al.* Harnessing the power of ChatGPT for automating systematic review process: methodology,

case study, limitations, and future directions. **Systems**, [*s. l.*], v. 11, n. 7, art. 351, 2023. Disponível em: https://doi.org/10.3390/systems11070351. Acesso em: 6 nov. 2026.

BANNACH-BROWN, Alexandra; PRIZYBYLA, Piotr; THOMAS, James; RICE, Andrew S. C. *et al.* Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error**. Systematic Reviews**, [*s. l.*], v. 8, art. 23, 2019. Disponível em: https://doi.org/10.1186/s13643-019-0942-7. Acesso em: 6 nov. 2026.

BORAH, Rohit; BROWN, Andrew W.; CAPERS, Patrice L.; KAISER, Kathryn A. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. **BMJ Open**, [*s. l.*], v. 7, n. 2, art. e012545, 2017. Disponível em: https://doi.org/10.1136/bmjopen-2016-012545. Acesso em: 6 nov. 2026.

BRAVERMAN, Harry. **Labor and monopoly capital**: the degradation of work in the twentieth century. New York: Monthly Review Press, 1974.

DAVENPORT, Thomas H.; KIRBY, Julia. **Only humans need apply**: winners and losers in the age of smart machines. New York: Harper Business, 2016.

GARTLEHNER, Gerald; KAHWATI, Leila; HILSCHER, Rainer; THOMAS, Ian *et al.* Data extraction for evidence synthesis using a large language model: a proof-of-concept study. **Research Synthesis Methods**, [*s. l.*], v. 15, n. 4, p. 576-589, 2024. Disponível em: https://doi.org/10.1002/jrsm.1710. Acesso em: 6 nov. 2026.

GUO, Eddie; GUPTA, Mehul; DENG, Jiawen; PARK, Ye-Jean *et al.* Automated paper screening for clinical reviews using large language models. **Journal of Medical Internet Research**, [*s. l.*], v. 26, art. e48996, 2024. Disponível em: https://doi.org/10.2196/48996. Acesso em: 6 nov. 2026.

HASAN, Bashar; SAADI, Samer; RAJJOUB, Noora S.; HEGAZI, Moustafa *et al.* Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. **BMJ Evidence-Based Medicine**, [*s. l.*], v. 29, n. 6, p. 394-398, 2024. Disponível em: https://doi.org/10.1136/bmjebm-2023-112597. Acesso em: 6 nov. 2026.

HIGGINS, Julian P. T.; THOMAS, James; CHANDLER, Jacqueline; CUMPSTON, Miranda *et al.* **Cochrane handbook for systematic reviews of interventions**. London: Cochrane, 2023. Disponível em: https://www.training.cochrane.org/handbook. Acesso em: 7 nov. 2026.

KHRAISHA, Qusai; PUT, Sophie; KAPPENBERG, Johanna; WARRAITCH, Azza *et al*. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. **Research Synthesis Methods**, [*s. l.*], v. 15, n. 4, p. 616-626, 2024. Disponível em: https://doi.org/10.1002/jrsm.1715. Acesso em: 6 nov. 2026.

LANDIS, J. Richard; KOCH, Gary G. The measurement of observer agreement for categorical data. **Biometrics**, [*s. l.*], v. 33, n. 1, p. 159-174, 1977. Disponível em: https://doi.org/10.2307/2529310. Acesso em: 6 nov. 2026.

LI, Michael; SUN, Jianping; TAN, Xianping. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. **Systematic Reviews**, [*s. l.*], v. 13, art. 219, 2024. Disponível em: https://doi.org/10.1186/s13643-024-02609-x. Acesso em: 6 nov. 2026.

MAHMOUDI, Hesam; CHANG, Doris; LEE, Hannah; GHAFFARZADEGAN, Navid *et al.* Critical assessment of large language models' (ChatGPT) performance in data extraction for systematic reviews: exploratory study. **JMIR AI**, [*s. l.*], v. 4, art. e68097, 2025. Disponível em: https://doi.org/10.2196/68097. Acesso em: 6 nov. 2026.

MARIANO, Ari Melo; ROCHA, Maíra Santos. Revisão da literatura: apresentação de uma abordagem integradora. *In*: AEDEM INTERNATIONAL CONFERENCE, 26., 2017, Reggio Calabria. **Anais** [...]. Reggio Calabria: AEDEM, 2017, p. 427-443. Disponível em: https://www.researchgate.net/publication/319547360_Revisao_da_Literatura_Apresentacao_de_uma_Abordagem_Integradora. Acesso em: 7 nov. 2026.

MARSHALL, Iain J.; WALLACE, Byron C. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. **Systematic Reviews**, [*s. l.*], v. 8, art. 163, 2019. Disponível em: https://doi.org/10.1186/s13643-019-1074-9. Acesso em: 6 nov. 2026.

MOTZFELDT JENSEN, Mette; BRIX DANIELSEN, Mathias; RIIS, Johannes; ASSIFUAH KRISTJANSEN, Karoline *et al.* ChatGPT-4o can serve as the second rater for data extraction in systematic reviews**. PLoS ONE**, [*s. l.*], v. 20, n. 1, art. e0313401, 2025. Disponível em: https://doi.org/10.1371/journal.pone.0313401. Acesso em: 6 nov. 2026.

NORDMANN, Kim; FISCHER, Florian. Capability of chatbots powered by large language models to support the screening process of scoping reviews: a feasibility study. **BMC Health Services Research**, [*s. l.*], v. 26, art. 109, 2026. Disponível em: https://doi.org/10.1186/s12913-025-13901-4. Acesso em: 6 nov. 2026.

OAMI, Takehiko; OKADA, Yohei; NAKADA, Taka-aki. GPT-3.5 Turbo and GPT-4 Turbo in title and abstract screening for systematic reviews. **JMIR Medical Informatics**, [*s. l.*], v. 13, art. e64682, 2025. Disponível em: https://doi.org/10.2196/64682. Acesso em: 6 nov. 2026.
OAMI, Takehiko; OKADA, Yohei; MAEDA, Kenjiro; NAKADA, Taka-aki. Performance of large language models and prompt engineering strategies for data extraction in systematic reviews. **Frontiers in Digital Health**, [*s. l.*], v. 8, art. 1799623, 2026. Disponível em: https://doi.org/10.3389/fdgth.2026.1799623. Acesso em: 6 nov. 2026.

O'MARA-EVES, Alison; THOMAS, James; McNAUGHT, John; MIWA, Makoto *et al*. Using text mining for study identification in systematic reviews: a systematic review of current approaches**. Systematic Reviews**, [*s. l.*], v. 4, art. 5, 2015. Disponível em: https://doi.org/10.1186/2046-4053-4-5. Acesso em: 6 nov. 2026.

PAGE, Matthew J.; McKENZIE, Joanne E.; BOSSUYT, Patrick M.; BOUTRON, Isabelle *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. **BMJ**, [*s. l.*],

v. 372, art. n71, 2021. Disponível em: https://doi.org/10.1136/bmj.n71. Acesso em: 6 nov. 2026.

REASON, Tim; BENBOW, Emma; LANGHAM, Julia; GIMBLETT, Andy *et al*. Artificial intelligence to automate network meta-analyses: four case studies to evaluate the potential application of large language models**. PharmacoEconomics** - Open, [*s. l.*], v. 8, n. 2, p. 205-220, 2024. Disponível em: https://doi.org/10.1007/s41669-024-00476-9. Acesso em: 6 nov. 2026.

SANGHERA, Rohan; HIRUNAVUKARASU, Arun James; EL KHOURY, Marc; O'LOGBON, Jessica *et al*. High-performance automated abstract screening with large language model ensembles**. Journal of the American Medical Informatics Association**, [*s. l.*], v. 32, n. 5, p. 893-904, 2025. Disponível em: https://doi.org/10.1093/jamia/ocaf050. Acesso em: 6 nov. 2026.

SOLLINI, Martina; PINI, Cristiano; LAZAR, Alexandra; GELARDI, Fabrizia; NINATTI, Gaia *et al*. Human researchers are superior to large language models in writing a medical systematic review in a comparative multitask assessment. **Scientific Reports**, [*s. l.*], v. 16, art. 173, 2025. Disponível em: https://doi.org/10.1038/s41598-025-28993-5. Acesso em: 6 nov. 2026.

SPILLIAS, Scott; OLLERHEAD, Katherine; ANDREOTTA, Matthew; ANNAND-JONES, Ruby *et al*. Evaluating generative AI for qualitative data extraction in community-based fisheries management literature**. Environmental Evidence***,*** [*s. l.*], v. 14, art. 9, 2025. Disponível em: https://doi.org/10.1186/s13750-025-00362-9. Acesso em: 6 nov. 2026.

SUSNJAK, Teo; HWANG, Peter; REYES, Napoleon H.; BARCZAK, Andre L. C. *et al*. Automating research synthesis with domain-specific large language model fine-tuning. **ACM Transactions on Knowledge Discovery from Data**, [*s. l.*], v. 19, n. 3, art. 68, 2025. Disponível em: https://doi.org/10.1145/3715964. Acesso em: 6 nov. 2026.

THIRUNAVUKARASU, Arun James; TING, Darren Shu Jeng; ELANGOVAN, Kabilan; GUTIERREZ, Laura *et al*. Large language models in medicine. **Nature Medicine**, [*s. l.*], v. 29, n. 8, p. 1930-1940, 2023. Disponível em: https://doi.org/10.1038/s41591-023-02448-8. Acesso em: 6 nov. 2026.

TRICCO, Andrea C.; LILLIE, Erin; ZARIN, Wasifa; O'BRIEN, Kelly K. *et al*. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation**. Annals of Internal Medicine**, [*s. l.*], v. 169, n. 7, p. 467-473, 2018. Disponível em: https://doi.org/10.7326/M18-0850. Acesso em: 6 nov. 2026.

VAN DIJK, Sanne H. B.; BRUSSE-KEIZER, Marjolein G. J.; BUCSÁN, Charlotte C.; VAN DER PALEN, Job *et al*. Artificial intelligence in systematic reviews: promising when appropriately used. **BMJ Open**, [*s. l.*], v. 13, n. 7, art. e072254, 2023. Disponível em: https://doi.org/10.1136/bmjopen-2023-072254. Acesso em: 6 nov. 2026.

VAN ECK, Nees Jan; WALTMAN, Ludo. Software survey: VOSviewer, a computer program for bibliometric mapping. **Scientometrics**, [*s. l.*], v. 84, n. 2, p. 523-538, 2010. Disponível em: https://doi.org/10.1007/s11192-009-0146-3. Acesso em: 6 nov. 2026.

YISHA, Zuhaer; OU, Peng; LI, Sheng; ZHANG, Lin *et al*. Large language models for data extraction in systematic reviews: a proof-of-concept study. **BMC Medical Research Methodology**, [*s. l.*], v. 26, art. 33, 2026. Disponível em: https://doi.org/10.1186/s12874-025-02729-5. Acesso em: 6 nov. 2026.

ZUBOFF, Shoshana. **The age of surveillance capitalism**: the fight for a human future at the new frontier of power. New York: PublicAffairs, 2019.