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

EFETIVIDADE DE UM MÉTODO ASSISTIDO POR LLMs COMO APOIO À TRIAGEM DE TRABALHOS EM REVISÃO DE LITERATURA

EFFECTIVENESS OF AN LLM-ASSISTED METHOD FOR SUPPORTING STUDY SCREENING IN LITERATURE REVIEWS

Renan Carneiro Batista – Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
Fabio Lorensi do Canto – Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
Washington Luís Ribeiro de Carvalho Segundo – Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
Thiago Magela Rodrigues Dias – Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)

Modalidade: Trabalho Completo

Resumo: Este trabalho avalia a efetividade de um método assistido por modelos de linguagem de grande escala (LLMs) para a triagem de trabalhos em procedimento simulado de revisão de literatura. O estudo foi conduzido a partir de um *corpus* de 69 registros recuperados no Portal Oasisbr a partir de uma busca booleana sobre o tema "uso de inteligência artificial no monitoramento do desmatamento e na preservação florestal na região amazônica". Os registros foram classificados por consenso pelos autores nas categorias Incluir, Reavaliar e Excluir, sendo a classificação adotada como referência para avaliar quatro modelos atuais de LLM, sendo dois proprietários (Gemini 3.1 Flash Lite e Claude Haiku 4.5) e dois abertos (Llama 3.1 8B e Qwen 3 14B). A avaliação considerou medidas de concordância, por meio do Kappa de Cohen, e de utilidade prática, por meio do *recall* da categoria Manter e do *Work Saved over Sampling at Recall R* (WSS@R). Os resultados indicam que Gemini, Claude e Qwen apresentaram concordância moderada com a classificação de referência humana, enquanto o Llama apresentou desempenho mais fraco e maior tendência de incluir trabalhos. O Gemini obteve o melhor resultado como pré-filtro, com *recall* de Manter de 100% e WSS@R de 31,9%. Conclui-se que LLMs podem apoiar a triagem preliminar e reduzir a carga de leitura, desde que utilizados sob supervisão humana.

Palavras-chave: recuperação da informação; revisão de literatura; inteligência artificial; modelos de linguagem.

Abstract: This study evaluates the effectiveness of a method assisted by large language models (LLMs) for screening studies in a simulated literature review procedure. The study was conducted using a corpus of 69 records retrieved from the Oasisbr Portal through a Boolean search on the topic "the use of artificial intelligence in monitoring deforestation and forest preservation in the Amazon region." The records were classified by consensus among the authors into three categories: Include, Reassess, and Exclude. This classification was adopted as the reference standard for evaluating four current LLMs: two proprietary models, Gemini 3.1 Flash Lite and Claude Haiku 4.5, and two open models, Llama 3.1 8B and Qwen 3 14B. The evaluation considered agreement measures, using Cohen's Kappa, and practical utility measures, using recall for the Keep category and Work Saved over Sampling at Recall R (WSS@R). The results indicate that Gemini, Claude, and Qwen showed moderate agreement with the human reference classification, whereas Llama showed weaker performance and a greater tendency to include studies. Gemini achieved the best result as a pre-filter, with 100% recall for the Keep category and a WSS@R of 31.9%. The study concludes that LLMs can support preliminary screening and reduce the reading workload, provided that they are used under human supervision.
Keywords: information retrieval; literature review; artificial intelligence; large language models.

1 INTRODUÇÃO

As revisões de literatura são elementos essenciais para a produção do conhecimento científico, atuando como instrumentos de síntese, sistematização e mapeamento do estado da arte de um tema de pesquisa. São procedimentos que sintetizam a produção, conectam contribuições dispersas sobre um mesmo assunto e ajudam na identificação de lacunas, tendências e possibilidades de avanço teórico e metodológico de um campo científico (Cooper, 1988, Webster; Watson, 2002).

Em muitos casos, a revisão de literatura compõe a fundamentação teórica do trabalho, servindo para contextualizar o problema, delimitar conceitos e justificar escolhas da pesquisa. Em outros casos, a revisão é a própria metodologia de coleta de dados, nos quais os documentos passam a ser o *corpus* analisado e a produção do conhecimento decorre da seleção, síntese, integração e crítica da literatura existente sobre um tema específico (Torraco, 2016; Snyder, 2019).

Nessa segunda perspectiva, as revisões podem assumir diferentes formatos, variando de abordagens narrativas, sistemáticas, integrativas, meta-análises e de escopo. Em algumas áreas, os resultados de estudos dessa natureza são considerados padrão ouro de evidência científica (Higgins; Thomas, 2024). Assim, a revisão deixa de ser apenas uma seção de apoio e

passa a operar como forma autônoma de pesquisa, capaz de gerar novas estruturas conceituais, perspectivas teóricas ou propostas de novos estudos (Grant; Booth, 2009; Snyder, 2019; Torraco, 2016).

O processo de revisar a literatura, contudo, é trabalhoso. A etapa de seleção ou triagem de trabalhos recuperados em bases de dados, na qual os pesquisadores lêem títulos, resumos e palavras-chave para decidir quais estudos avançam para a leitura de texto completo, consome um tempo considerável dos pesquisadores (Huotala; Kuutila; Ralph; Mantyla, 2024). Isso se agrava ainda mais à medida que cresce o volume de produção científica mundial (Harari Parola; Hartwell; Riegelman, 2020; Polanin; Pigott; Espelage; Grotpeter, 2019).

Mas o avanço dos *Large Language Models* (LLMs) oferece a possibilidade de automação desse processo de avaliação de trabalhos com base em critérios de inclusão e exclusão previamente definidos. Os modelos atuais são capazes de interpretar textos científicos a partir de comandos formulados em linguagem natural, dispensando pré-treinamento específico para esta atividade (Behrouzian *et al.*, 2026; Huotala; Kuutila; Ralph; Mantyla, 2024; Li; Sun; Tan, 2024; Shameem *et al.*, 2026).

Estudos recentes demonstram o potencial de uso dessas ferramentas como apoio do processo de revisão de literatura. Li, Sun e Tan (2024) apontam que o uso de APIs de LLMs reduz drasticamente o tempo e o custo financeiro da triagem quando comparado ao trabalho humano tradicional. Huotala, Kuutila, Ralph e Mantyla (2024) ressaltam que, embora as ferramentas automatizadas ainda necessitem de supervisão humana para mitigar o risco de exclusões incorretas, elas aceleram o processo ao direcionar a atenção dos pesquisadores para os estudos mais relevantes, reduzindo o volume de texto que precisa ser lido para tomada de decisão.

Mas a escolha do modelo e a estruturação das instruções são fatores relevantes para o resultado. Modelos proprietários, como o *ChatGPT* v4.0, alcançam acurácia superior a 90%, equiparando-se ao padrão de especialistas humanos (Li; Sun; Tan, 2024). O uso de técnicas de engenharia de *prompt*, como o *Chain-of-Thought*, reduz significativamente os erros de classificação quando comparado a instruções diretas (Huotala; Kuutila; Ralph; Mantyla, 2024).

Além disso, o desempenho dos modelos é afetado não apenas por sua capacidade semântica, mas também pela forma como os critérios de elegibilidade são redigidos e pela arquitetura de decisão adotada. Em *prompts* rígidos e sequenciais, nos quais o não

atendimento de um critério leva à exclusão imediata do registro, ambiguidades ou informações ausentes no resumo podem resultar em exclusões indevidas e, consequentemente, em perda de *recall* (Behrouzian *et al*., 2026).

Mas os estudos são unânimes na premissa de que as LLMs ainda não substituem totalmente o julgamento humano. Em uma revisão de literatura, as decisões de inclusão e exclusão podem comprometer o rigor metodológico do estudo. O uso de LLMs se coloca, portanto, como apoio à triagem, um pré-filtro sujeito à supervisão do pesquisador, não atuando em sua substituição (Behrouzian *et al*., 2026; Huotala; Kuutila; Ralph; Mantyla, 2024; Li; Sun; Tan, 2024; Shameem *et al.*, 2026).

Diante dessa nova perspectiva, este trabalho tem por objetivo avaliar a efetividade de um método assistido por LLMs como apoio à etapa de triagem de uma revisão de literatura a partir de um tema de pesquisa regional e com estudos majoritariamente em língua portuguesa. Efetividade, neste caso, é definida como a capacidade dos modelos de reduzir a carga de leitura sem perda de estudos considerados relevantes, tendo como referência uma classificação definida por consenso pelos autores.

Complementarmente, o estudo busca comparar o desempenho entre os modelos comerciais e de código aberto, mensurar a concordância entre os modelos e a classificação de referência e estimar a economia de trabalho de leitura proporcionada pelo método assistido, sem perda de estudos classificados como relevantes.

Apesar de seu caráter preliminar e delimitado a um cenário de triagem simulada, este estudo poderá fornecer subsídios para o desenvolvimento e a validação de modelos de LLM aplicados à síntese do conhecimento em procedimentos metodológicos de revisão de literatura.

2 PROCEDIMENTOS METODOLÓGICOS

O tema definido para o procedimento simulado de triagem de trabalhos para uma revisão de literatura é o uso de Inteligência Artificial (IA) aplicada ao monitoramento de desmatamento e à preservação florestal na região amazônica.

2.1 CONSTRUÇÃO DO *CORPUS*

Para construção do *corpus* a partir do tema foi realizada uma busca no Portal Oasisbr[1]

[1] https://oasisbr.ibict.br/.

com a seguinte expressão booleana: (“inteligência artificial” *OR* “aprendizado de máquina” *OR* “aprendizado profundo” *OR* “machine learning” *OR* “deep learning” *OR* “modelos de linguagem”) *AND* (desmatamento *OR* deflorestamento *OR* “monitoramento florestal” *OR* “conservação da biodiversidade” *OR* “conservação florestal”) *AND* (Amazônia *OR* Amazonia *OR* “região amazônica” *OR* “floresta amazônica” *OR* “bacia amazônica”).

A busca foi realizada em maio de 2026, sem recorte temporal, abrangendo todo o período coberto pelo Portal. Dos 71 resultados apresentados na interface de busca, a exportação em formato RIS retornou 70 registros, um dos quais truncado no próprio arquivo de exportação (registro inválido, do tipo notícia institucional, descartado na leitura automatizada). O corpus final reuniu, portanto, 69 registros. Optou-se por não realizar limpeza prévia dos metadados: um registro sem resumo e treze com resumos fora do padrão descritivo, contendo notas administrativas como a identificação do orientador ou a descrição do tipo de documento, foram deliberadamente mantidos, de modo a avaliar o comportamento dos modelos diante de metadados imperfeitos, condição recorrente em bases de dados reais.

2.2 CRITÉRIOS DE ELEGIBILIDADE

Foram definidos três critérios obrigatórios para inclusão dos trabalhos selecionados no *corpus*: 1) empregar técnicas de IA, aprendizado de máquina, aprendizado profundo, visão computacional ou processamento de linguagem natural, incluindo a aplicação de redes neurais a imagens de sensoriamento remoto; 2) estudos sobre desmatamento, conservação ou monitoramento florestal, clima ou biodiversidade; 3) e ser aplicado à Amazônia em sentido pan-amazônico, ou seja, abrangendo a região amazônica em qualquer dos países nos quais o bioma é considerado.

Como critério de exclusão, foram desconsiderados os registros que não atenderam a qualquer um dos critérios obrigatórios, como estudos sem uso de técnica de IA, sem perspectiva de preservação florestal ou realizados em outros biomas. Registros com aderência parcial ou resumo sem informações suficientes foram classificados em uma categoria intermediária (Reavaliar), que indica a necessidade de leitura do texto completo para avaliação definitiva.

2.3 PROCEDIMENTOS DE TRIAGEM

Os dois avaliadores humanos e os modelos de linguagem receberam os mesmos 69 registros e aplicaram os mesmos critérios, classificando cada um dos registros de acordo com as categorias Incluir, Reavaliar ou Excluir.

2.3.1 Triagem humana

A triagem humana de referência foi realizada pelos autores deste trabalho. Primeiramente, foi realizada uma triagem por dois avaliadores de forma independente e sob o método de análise cega, assegurando que as avaliações preliminares ocorressem sem interferência mútua. Em uma etapa subsequente, as divergências identificadas foram submetidas ao debate para busca de consenso, recorrendo-se a um terceiro avaliador responsável pelo parecer final de desempate. A classificação final resultante foi adotada como o padrão de referência para avaliar o método assistido.

2.3.2 Método assistido por LLMs

Foram avaliados quatro modelos de LLM (Quadro 1), representando dois paradigmas de acesso: proprietários de maior capacidade, acessados por interface remota, e abertos, executáveis localmente em um ambiente do *Google Colab* configurado com uma GPU Nvidia T4 de 16 GB.

Quadro 1 - Modelos de linguagem avaliados

Modelo	Provedor	Regime	Acesso
Gemini 3.1 Flash Lite	Google	Comercial	Remoto
Claude Haiku 4.5	Anthropic	Comercial	Remoto
Llama 3.1 8B	Meta	Aberto	Local
Qwen 3 14B	Alibaba	Aberto	Local

Fonte: elaboração própria (2026).

Cada registro foi submetido a todos os modelos com o mesmo *prompt* e esquema de saída, garantindo uma justa comparação. O *prompt* orientou os modelos a raciocinar em três fases: analisar cada pilar com base nos metadados (título, resumo e palavras-chave), sintetizar pontos favoráveis e contrários à inclusão e, por fim, classificar o registro com um indicador de confiança e uma justificativa.

A saída segue um esquema estruturado de doze campos (Quadro 2), no qual a decisão é posicionada após os campos de análise para induzir o modelo a produzir evidências intermediárias antes da decisão final, técnica conhecida como *chain-of-thought* (Wei *et al*.,

2022).

Quadro 2 - Esquema estruturado de saída do modelo

Campo	Tipo de conteúdo
Análise de tecnologia	Texto livre
Análise por domínio	Texto livre
Análise de geografia	Texto livre
Tecnologia atendida	Booleano
Domínio atendido	Booleano
Geografia atendida	Booleano
Pontos a favor da inclusão	Texto livre
Pontos contra a inclusão	Texto livre
Informação suficiente para decisão	Booleano
Decisão	Incluir / Reavaliar / Excluir
Confiança na decisão	Numérico em [0, 1]
Justificativa final	Texto livre

Fonte: elaboração própria (2026).

O *prompt* incorpora exemplos de calibragem para cada categoria utilizando a técnica de *few-shot prompting* (Brown *et al*., 2020). Eventuais saídas inválidas foram tratadas por validação automática e novas tentativas; registros que permaneceram inválidos foram contabilizados como falhas, sem substituição por estimativa. Todos os modelos foram executados com temperatura 0, minimizando a variabilidade das saídas.

2.4 MÉTRICAS DE AVALIAÇÃO

A avaliação articula duas perspectivas complementares. A primeira mede o grau de acordo entre cada modelo e a classificação de referência por meio do coeficiente Kappa de Cohen, adotado por corrigir a concordância esperada ao acaso, correção relevante em um corpus desbalanceado como o deste estudo, no qual a concordância bruta tende a superestimar o acordo. São reportadas duas variantes: o *Kappa* simples, que trata as categorias como nominais, e o *Kappa* ponderado com pesos quadráticos, mais adequado à natureza ordinal das categorias (Excluir < Reavaliar < Incluir) por penalizar as discordâncias proporcionalmente à distância entre elas. Adota-se a interpretação das faixas de Landis e Koch (1977). Dado o tamanho reduzido do corpus, os intervalos de confiança de 95% do Kappa quadrático foram estimados por *bootstrap* não paramétrico, com 2.000 reamostragens dos registros e cálculo pelo método dos percentis.

A utilidade prática como pré-filtro mede a economia de leitura. Para isso, as categorias

são agrupadas em Descartar (Excluir) e Manter (Reavaliar e Incluir), e calcula-se o *Work Saved over Sampling at Recall R* (WSS@R) de Cohen, Hersh, Peterson, Yen (2006), definido como:

$$WSS@R = (TN + FN) / N - 1\ (1 - R) \qquad (1)$$

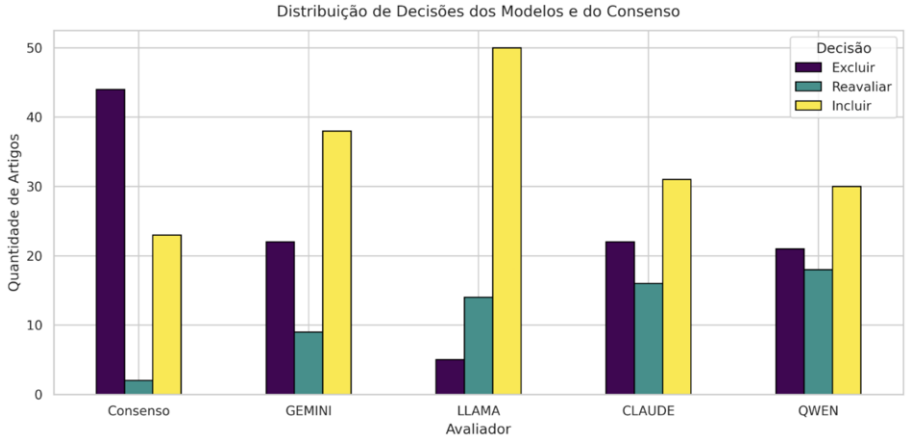
Em que TN e FN são os descartes corretos e incorretos, N o total de registros e R o *recall* da categoria Manter. O *recall* de Manter é condição não negociável, pois um WSS elevado obtido às custas de falsos descartes implicaria em perda de literatura relevante.

O delineamento adotado caracteriza-se como um estudo de caso: um único tema, um único *corpus* e uma única execução por modelo. Nesse desenho, a generalização pretendida é analítica, e não estatística: o que se propõe como transferível não são os valores obtidos pelos modelos, mas o protocolo de avaliação (critérios explícitos, triagem humana de consenso como referência e métricas de concordância e de economia de leitura), que pode ser replicado em outros corpora e domínios.

3 RESULTADOS E DISCUSSÃO

Os registros do *corpus* foram submetidos aos quatro modelos sob os mesmos critérios e o mesmo esquema de saída adotados na triagem humana. Os quatro modelos produziram classificação válida para a totalidade dos registros, sem falhas residuais após a validação automática. As decisões foram avaliadas em relação à classificação de referência.

Gráfico 1 - Distribuição das decisões dos modelos e da referência



Fonte: elaboração própria (2026).

O Gráfico 1 apresenta a distribuição das decisões dos modelos frente à referência.

3.1 CONCORDÂNCIA ENTRE MODELOS E A REFERÊNCIA

As decisões dos modelos distribuíram-se de modo distinto. O Gemini 3.1 Flash-Lite

classificou 22 registros como Excluir, 9 como Reavaliar e 38 como Incluir. O Claude Haiku 4.5 classificou 22 como Excluir, 16 como Reavaliar e 31 como Incluir. O Qwen 3 14B classificou 21 como Excluir, 18 como Reavaliar e 30 como Incluir. O Llama 3.1 8B classificou 5 como Excluir, 14 como Reavaliar e 50 como Incluir.

A Tabela 1 reúne as medidas de concordância de cada modelo com a classificação de referência. O intervalo de confiança de 95% foi estimado por reamostragem.

Tabela 1 - Concordância modelos e classificação de referência

Modelo	Concordância Bruta (%)	Kappa Simples	Kappa quad.	IC 95% (quad.)	*Recall* Incluir (%)
Gemini 3.1 Flash Lite	63,8%	0,405	0,510	[0,35; 0,67]	95,7%
Claude Haiku 4.5	59,4%	0,366	0,538	[0,37; 0,70]	87,0%
Qwen 3 14B	52,2%	0,268	0,417	[0,24; 0,59]	69,6%
Llama 3.1 8B	37,7%	0,118	0,130	[0,02; 0,25]	87,0%

Fonte: elaboração própria (2026).

Com base em Landis e Koch (1977) sobre o kappa quadrático, Gemini, Claude e Qwen situam-se na faixa moderada, e Llama permanece na faixa fraca. Os intervalos de confiança dos três primeiros se sobrepõem entre si e não se sobrepõem ao do Llama, cujo limite superior fica abaixo dos limites inferiores dos demais. Registra-se ainda um padrão de hesitação, com os modelos alocando na categoria Reavaliar parte dos registros definidos como Excluir ou Incluir na classificação de referência.

O exame da distribuição dos erros esclarece a origem dos níveis de concordância observados. Dos 69 registros, a classificação de referência sinaliza 44 como Excluir e 25 como Manter, reunindo os 23 de Incluir e os 2 de Reavaliar, o que torna o *corpus* desbalanceado em favor da exclusão. A discordância entre os modelos e a referência concentra-se quase toda nos registros classificados como Excluir, e não nos classificados como Incluir. Os quatro modelos foram capazes de recuperar a categoria Incluir e raramente cometeram o erro de descartar um registro relevante. O erro predominante foi o oposto, a retenção indevida de registros que o consenso havia excluído, ora alocados em Reavaliar, ora promovidos a Incluir. A Tabela 2 detalha o destino dos 44 registros que o consenso classificou como Excluir.

Tabela 2 - Destino dos registros classificados como Excluir no consenso (n = 44)

Modelo	Mantidos em Excluir	Movidos para Reavaliar	Movidos para Incluir
Gemini 3.1 Flash Lite	22	8	14
Claude Haiku 4.5	21	14	9
Qwen 3 14B	20	12	12

Llama 3.1 8B	5	10	29

Fonte: elaboração própria (2026).

O quadro revela gradações de uma mesma tendência inclusiva. Gemini, Claude e Qwen mantiveram em Excluir cerca de metade dos registros que deveriam ser descartados, entre 20 e 22 dos 44, e distribuíram os demais entre Reavaliar e Incluir. O Llama destoa do conjunto, pois manteve apenas cinco registros em Excluir e classificou 29 como Incluir. Esse comportamento confirma sua tendência inclusiva e explica seu desempenho fraco nas medidas de concordância.

A leitura conjunta da Tabela 1 mostra ainda que o *recall* de Incluir e a concordância geral não acompanham um ao outro. O Qwen obteve *recall* de Incluir de 69,6%, abaixo dos 87,0% do Claude e Llama e dos 95,7% do Gemini, embora seu kappa não seja o mais baixo da tabela. A diferença decorre do uso da categoria intermediária, já que o Qwen direcionou para Reavaliar parte dos registros que o consenso considerou Incluir. Esses registros não contam como acerto no *recall* de Incluir, mas permanecem retidos quando as categorias são agrupadas em Manter, o que desloca a posição relativa do modelo conforme a métrica adotada. A escolha da medida, portanto, não é neutra, e define qual modelo parece melhor.

Cabe uma ressalva sobre a própria referência. Na triagem cega, os dois avaliadores humanos recorreram à categoria Reavaliar em 29 dos 69 registros, mas a etapa de consenso e o desempate resolveram quase todas essas indecisões, de modo que a classificação final reteve Reavaliar em apenas dois registros. Os modelos produziram um único parecer, sem etapa de deliberação, e alocaram em Reavaliar entre 9 e 18 registros. Parte da hesitação atribuída aos modelos reflete essa diferença de procedimento, pois a referência passou por uma resolução que os modelos não tiveram. Essa assimetria reduz a concordância simples e reforça a cautela ao interpretar o kappa como medida isolada de qualidade.

A concordância entre os próprios modelos completa o quadro. Calculada sobre a decisão binária de manter ou descartar, ela alcançou cerca de 90% entre Gemini, Claude e Qwen, e ficou em torno de 75% entre o Llama e os demais. O Llama destaca-se como ponto fora da curva mesmo sem referência humana, o que aumenta a confiança na sua identificação como o modelo de comportamento mais discrepante.

3.2 ECONOMIA DE LEITURA

Para avaliar a utilidade prática do método como pré-filtro, as três categorias foram

agrupadas em Descartar, que reúne os registros classificados como Excluir, e Manter, que reúne Reavaliar e Incluir. O critério que organiza essa avaliação é assimétrico. Na triagem para revisão de literatura, descartar um estudo relevante custa mais do que reter um irrelevante, porque uma exclusão indevida reduz a cobertura da revisão e não é recuperada nas etapas seguintes. Por isso, a medida que condiciona a utilidade do filtro é o *recall* da categoria Manter, e não a acurácia ou o kappa. O WSS@R, proposto por Cohen, Hersh, Peterson, Yen (2006), foi formulado sobre essa premissa, ao fixar um patamar de *recall* e medir a economia de leitura obtida sem o ultrapassar.

A economia de leitura tem duas leituras complementares. A primeira é a redução bruta de trabalho, dada pela fração de registros que o filtro retira da leitura de texto completo. Gemini e Claude descartaram 22 dos 69 registros cada, cerca de 32%, o Qwen descartou 21, cerca de 30%, e o Llama descartou 5, cerca de 7%. A segunda leitura é o WSS@R, que desconta dessa redução o custo dos registros relevantes perdidos. Quando o *recall* de Manter é de 100%, as duas medidas coincidem, como no Gemini, que manteve a redução de 31,9% sem nenhum descarte incorreto. Quando há perda, a diferença entre as duas medidas é o preço dessa perda. Claude e Qwen descartaram volume semelhante ao do Gemini, porém com um descarte incorreto cada, o que reduziu o WSS para 27,9% e 26,4%.

O Llama ilustra o limite oposto. Preservou todo o material relevante, mas descartou apenas cinco registros, o que reflete sua classificação majoritariamente inclusiva e resulta em economia de 7,2%. Cohen, Hersh, Peterson, Yen (2006) relataram redução de trabalho em onze de quinze revisões analisadas, com reduções de 50% ou mais em três delas. Os valores próximos de 30% obtidos por Gemini, Claude e Qwen ficam abaixo desses melhores casos, mas representam economia real de leitura preservando o material relevante, que é o resultado esperado de um pré-filtro.

A escolha do patamar de *recall* define quais modelos são aceitáveis. Sob exigência de 100% de *recall* para Manter, condição mais conservadora, apenas Gemini e Llama a cumprem. Sob exigência de 95%, patamar adotado por Cohen, Hersh, Peterson, Yen (2006), os quatro a cumprem. Essa escolha não é técnica, e sim metodológica, pois depende de quanto risco de perda o pesquisador aceita em troca de economia de leitura.

A ordenação dos modelos muda conforme a medida considerada. Pelo kappa quadrático da Tabela 1, o Claude lidera com 0,538, seguido por Gemini com 0,510, Qwen com 0,417 e Llama com 0,130. Pelo WSS@R da Tabela 3, a ordem se inverte. O Gemini passa a

primeiro com 31,9%, o Claude a segundo com 27,9%, e Qwen e Llama seguem com 26,4% e 7,2%.

Tabela 3 - Utilidade do pré-filtro em relação à referência

Modelo	Registros descartados	Descartes incorretos	*Recall* Manter (%)	WSS@R
Gemini 3.1 Flash Lite	22	0	100%	31,9
Claude Haiku 4.5	22	1	96%	27,9
Qwen 3 14B	21	1	96%	26,4
Llama 3.1 8B	5	0	100%	7,2

Fonte: elaboração própria (2026).

As duas medidas não recompensam a mesma propriedade. O kappa avalia a concordância ordinal nas três categorias e penaliza o trânsito de registros pela Reavaliar. O WSS@R mede quanto de leitura se economiza ao descartar registros sem violar o *recall* de Manter. Um modelo pode então concordar menos com a referência categoria a categoria e ainda descartar uma fração maior de registros irrelevantes no patamar de *recall* adotado. O Gemini ilustra esse comportamento, pois filtra melhor sob o critério de economia mesmo concordando menos no detalhe. Não há um modelo superior em termos absolutos. A escolha depende do objetivo, se concordância fiel à referência ou economia de leitura como pré-filtro, e o Llama fica em último sob as duas medidas.

3.3 SÍNTESE DOS RESULTADOS

Um dos objetivos do estudo era comparar modelos abertos e proprietários. Os resultados não revelam uma fronteira nítida entre as duas categorias. O Qwen, aberto e com 14 bilhões de parâmetros, acompanhou os modelos comerciais nas medidas de concordância e de economia de leitura. O Llama, também aberto, mas com 8 bilhões de parâmetros, foi o destoante em todas as medidas. A diferença observada parece responder mais à escala e ao ajuste do modelo do que à licença de uso. Como apenas dois modelos abertos foram avaliados, não é possível isolar a causa. Ainda assim, o desempenho do Qwen sugere que modelos abertos de porte médio são viáveis como pré-filtro sem custo de API.

De forma exploratória, examinou-se se a confiança declarada pelos modelos poderia servir de sinal para priorizar a revisão humana. Apenas o Claude apresentou separação relevante entre a confiança nos acertos, de 0,94, e nos erros, de 0,77. Nos demais modelos a diferença foi desprezível, e todos relataram confiança alta mesmo quando erravam. O número de erros por modelo é pequeno, o que limita a leitura desse resultado. Com essa ressalva, a

confiança auto atribuída não se mostrou um sinal confiável para triagem.

O estudo usou um único *corpus* de 69 registros, de um domínio específico, e cada modelo foi executado uma única vez, o que impede medir a estabilidade das classificações entre execuções. A referência foi construída por dois avaliadores, número reduzido para estimar a concordância humana. Os intervalos de confiança são largos e se sobrepõem entre os modelos, de modo que as diferenças observadas indicam tendências, e não ordenações definitivas.

Como resultado adicional, o estudo disponibiliza publicamente o conjunto completo de materiais (o *corpus*, os *prompts*, as classificações individuais e de consenso dos avaliadores e as saídas integrais dos modelos), que constitui um padrão de referência reutilizável para a avaliação de outros modelos e estratégias de triagem sobre o mesmo tema[2].

No conjunto, os achados sustentam a viabilidade do método como pré-filtro assistido, e não como substituto da triagem humana. O erro dominante dos modelos foi a inclusão indevida, e não o descarte de registros relevantes, o que é o modo de falha preferível em revisões de literatura. A escolha do modelo e do patamar de *recall* permanece uma decisão metodológica, condicionada ao risco de perda que o pesquisador aceita assumir em troca de economia de leitura.

4 CONSIDERAÇÕES FINAIS

Esses achados sugerem que LLMs podem contribuir para reduzir a carga de leitura em etapas iniciais de triagem, especialmente quando utilizados para separar registros claramente irrelevantes daqueles que devem seguir para avaliação humana. No entanto, os resultados também reforçam que os modelos ainda não têm capacidade de substituir o julgamento humano. Em revisões de literatura, a exclusão indevida de registros pode comprometer a validade do *corpus* e, consequentemente, a qualidade da síntese produzida. Por isso, o uso de LLMs ainda deve ser compreendido como apoio metodológico supervisionado, e não como mecanismo autônomo de decisão.

Dado seu caráter preliminar, o estudo apresenta limitações importantes. Foi utilizado um único *corpus*, com número reduzido de registros e de um tema único. Além disso, os resultados dependem do *prompt* utilizado, dos modelos selecionados, da qualidade dos

[2] Disponível em: https://github.com/Renan-Carneiro-Batista/triagem-llm. Acesso em: 12 set. 2026.

resumos e da classificação de referência adotada.

Como pesquisas futuras, pretende-se ampliar o *corpus*, testar temas de diversas áreas, comparar diferentes formulações de *prompt* e utilizar outras formas de avaliação dos modelos. A utilização de um *benchmark* mais amplo de triagens produzidas por avaliadores independentes também poderá fortalecer a validade da classificação de referência e permitir uma avaliação mais robusta dos resultados.

Ainda assim, os resultados preliminares indicam que métodos assistidos por LLMs podem ser promissores para apoiar revisões de literatura, desde que orientados por critérios explícitos, métricas de segurança e supervisão humana.

REFERÊNCIAS

BEHROUZIAN, Amir M. *et al.* Linguistic polarity and decision architecture in LLM-based abstract screening for systematic reviews. **Information**, [*s. l.*], v. 17, n. 5, art. 449, maio 2026. Disponível em: https://doi.org/10.3390/info17050449. Acesso em: 21 jun. 2026.

BROWN, Tom B.; MANN, Benjamin; RYDER, Nick; SUBBIAH, Melanie *et al.* Language models are few-shot learners. **arXiv**, [s*. l.*], arXiv:2005.14165, 2020. Disponível em: https://arxiv.org/abs/2005.14165. Acesso em: 21 jun. 2026.

COHEN, Aaron M.; HERSH, William R.; PETERSON, K.; YEN, Po-Yin *et al.* Reducing workload in systematic review preparation using automated citation classification. **Journal of the American Medical Informatics Association**, [*s. l.*], v. 13, n. 2, p. 206-219, 2006. Disponível em: https://doi.org/10.1197/jamia.M1929. Acesso em: 21 jun. 2026.

COOPER, Harris M. Organizing knowledge syntheses: a taxonomy of literature reviews. **Knowledge in Society**, [*s. l.*], v. 1, n. 1, p. 104-126, 1988. Disponível em: https://doi.org/10.1007/BF03177550. Acesso em: 16 jun. 2026.

GRANT, Maria J.; BOOTH, Andrew. A typology of reviews: an analysis of 14 review types and associated methodologies. **Health Information & Libraries Journal**, [*s. l.*], v. 26, n. 2, p. 91-108, 2009. Disponível em: https://doi.org/10.1111/j.1471-1842.2009.00848.x. Acesso em: 16 jun. 2026.

HARARI, Michael B.; PAROLA, Heather R.; HARTWELL, Christopher J.; RIEGELMAN *et al.* Literature searches in systematic reviews and meta-analyses: a review, evaluation, and recommendations. **Journal of Vocational Behavior**, [*s. l.*], v. 118, art. 103377, 2020. Disponível em: https://doi.org/10.1016/j.jvb.2020.103377. Acesso em: 16 jun. 2026.

HIGGINS, Julian P. T.; THOMAS, James; CHANDLER, Jacqueline; CUMPSTON, Miranda *et al.* (ed.). **Cochrane handbook for systematic reviews of interventions**. [*S. l.*]: Cochrane, 2024. Disponível em: https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current. Acesso em: 20 jun. 2026.

HUOTALA, Aleksi; KUUTILA, Miikka; RALPH, Paul; MÄNTYLÄ, Mika. The promise and challenges of using LLMs to accelerate the screening process of systematic reviews. *In*: INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING, 28., 2024, Salerno. **Anais** [...]. New York: ACM, 2024. p. 262-271. Disponível em: https://doi.org/10.1145/3661167.3661172. Acesso em: 18 jun. 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: 21 jun. 2026.

LI, Michael; SUN, Jianping; TAN, Xianming. 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: 18 jun. 2026.

POLANIN, Joshua R.; PIGOTT, Therese D.; ESPELAGE, Dorothy L.; GROTPETER, Jennifer K. Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. **Research Synthesis Methods**, [*s. l.*], v. 10, n. 3, p. 330-342, 2019. Disponível em: https://doi.org/10.1002/jrsm.1354. Acesso em: 16 jun. 2026.

SHAMEEM, Mohammad; ALSHAYEB, Mohammad; NIAZI, Mahmood; MAHMOOD, Sajjad; AHMAD, Aakash; AL DALLAL, Jehad. An exploratory study on effectiveness of GPT-4o in conducting sub-tasks of systematic literature reviews. **Automated Software Engineering**, [*s. l.*], v. 33, n. 3, art. 81, 2026. Disponível em: https://doi.org/10.1007/s10515-026-00629-6. Acesso em: 21 jun. 2026.

SNYDER, Hannah. Literature review as a research methodology: an overview and guidelines. **Journal of Business Research**, [*s. l.*], v. 104, p. 333-339, 2019. Disponível em: https://doi.org/10.1016/j.jbusres.2019.07.039. Acesso em: 16 jun. 2026.

TORRACO, Richard J. Writing integrative literature reviews: using the past and present to explore the future. **Human Resource Development Review**, [*s. l.*], v. 15, n. 4, p. 404-428, 2016. Disponível em: https://doi.org/10.1177/1534484316671606. Acesso em: 16 jun. 2026.

WEBSTER, Jane; WATSON, Richard T. Analyzing the past to prepare for the future: writing a literature review. **MIS Quarterly**, [*s. l.*], v. 26, n. 2, p. xiii-xxiii, 2002. Disponível em: https://www.jstor.org/stable/4132319. Acesso em: 16 jun. 2026.
WEI, Jason; WANG, Xuezhi; SCHUURMANS, Dale; BOSMA, Maarten *et al.* **Chain-of-thought prompting elicits reasoning in large language models**. arXiv, [*s. l.*], arXiv:2201.11903, 2022. Disponível em: https://arxiv.org/abs/2201.11903. Acesso em: 21 jun. 2026.