v7.n12.2026

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

PRÁTICAS DE AVALIAÇÃO DE PRODUTOS DE *SOFTWARE* BASEADOS EM LLM AO LONGO DO CICLO DE VIDA: UMA REVISÃO DE ESCOPO

EVALUATION PRACTICES FOR LLM-BASED SOFTWARE PRODUCTS ACROSS THE LIFECYCLE: A SCOPING REVIEW

PRÁCTICAS DE EVALUACIÓN DE PRODUCTOS DE SOFTWARE BASADOS EN LLM A LO LARGO DEL CICLO DE VIDA: UNA REVISIÓN DE ALCANCE

Luis Gustavo Olimpio Campos da Silva[1], Carlos Futino Barreto[2]

https://doi.org/10.47820/recima21.v7i12.10037

PUBLICADO: 12/2026

RESUMO

Produtos de *software* baseados em modelo de linguagem de grande porte (LLM) têm comportamento não determinístico, o que torna insuficientes os testes tradicionais. Desde 2022, indústria e academia construíram práticas de avaliação, dispersas entre engenharia de *software*, interação humano-computador e processamento de linguagem natural. Este trabalho mapeia essa avaliação ao longo do ciclo de vida, por revisão de escopo (PRISMA-ScR) com protocolo registrado. A busca em três braços (OpenAlex, DBLP e *snowballing*), em agosto de 2026, identificou 2.116 registros (1.681 após deduplicação). A triagem usou instrumento baseado em LLM, validado contra 393 decisões do autor (revocação 90,8%, precisão 70,1%), com auditoria humana e segunda pessoa revisora (*kappa* 0,50). O mapa de 881 estudos, classificados por título e resumo com instrumento validado contra 40 classificações cegas, mostra avaliação conduzida por pesquisadores (67,4%), em cenário acadêmico e antes da produção, a fase menos documentada (16,7%); métodos baseados em dados de uso ou de execução aparecem em 7,7% e especialistas de domínio participam em 21,0%. A leitura integral de 46 estudos centrais organiza os desafios em dez temas: a dificuldade começa na definição do que é correto e se repete em cada instrumento que tenta substituí-la. Na própria condução, nenhum modelo isolado, nem a dupla humana (*kappa* 0,50), atingiu a concordância de referência para triagem por revisor único; o teste-reteste do autor deu *kappa* 0,57. Todos divergiram do autor na mesma direção, excluindo casos-limite, e a calibração com exemplos esteve associada a maior concordância, em amostra que sobrerrepresenta os casos-limite.

PALAVRAS-CHAVE: Avaliação de sistemas baseados em LLM. Revisão de escopo. Engenharia de *software*.

ABSTRACT

Software products whose core capability is a large language model (LLM) behave nondeterministically, making conventional testing insufficient. Since 2022, industry and academia have built evaluation practices for these products, documented in fragments across software engineering, human-computer interaction, and natural language processing. This work maps how such evaluation is practiced across the lifecycle through a scoping review (PRISMA-ScR) with a registered protocol. A three-arm hybrid search (OpenAlex, DBLP, and snowballing), run in August 2026, identified 2,116 records (1,681 after deduplication). Screening used an LLM-based instrument validated on 393 author decisions (recall 90.8%, precision 70.1%), with human audit and a second reviewer (kappa 0.50). The map of 881 studies, charted from titles and abstracts with an instrument validated against 40 blind classifications, shows evaluation conducted mostly by researchers (67.4%), in academic settings and before production; production is the least documented phase (16.7%), methods based on usage or execution data appear in 7.7% of

[1] Graduando em Engenharia de *Software* pelo iCEV - Instituto de Ensino Superior, Teresina, PI, Brasil.
[2] Mestre em Engenharia de *Software* pelo Centro de Estudos e Sistemas Avançados do Recife (CESAR). Professor do curso de Engenharia de *Software* do iCEV - Instituto de Ensino Superior, Teresina, PI, Brasil.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

studies, and domain experts take part in 21.0%. Full-text reading of 46 central studies organizes challenges into ten themes: the difficulty starts with defining what a correct answer is and recurs in every instrument meant to replace that definition. In the review itself, no individual model, nor the human pair (kappa 0.50), reached the reference agreement level for single-reviewer screening; the author's test-retest yielded kappa 0.57. All diverged from the author in the same direction, excluding borderline cases, and calibration with examples was associated with higher agreement on a sample that over-represents borderline cases.

KEYWORDS: *Evaluation of LLM-based systems. Scoping review. Software engineering.*

RESUMEN

Los productos de software basados en modelos de lenguaje de gran tamaño (LLM) presentan comportamiento no determinista, lo que hace insuficientes las pruebas tradicionales. Desde 2022, industria y academia han construido prácticas de evaluación, dispersas entre ingeniería de software, interacción humano-computadora y procesamiento del lenguaje natural. Este trabajo mapea dicha evaluación a lo largo del ciclo de vida mediante revisión de alcance (PRISMA-ScR) con protocolo registrado. Una búsqueda en tres brazos (OpenAlex, DBLP y snowballing), en agosto de 2026, identificó 2.116 registros (1.681 tras deduplicar). El cribado utilizó instrumento basado en LLM, validado en 393 decisiones del autor (exhaustividad 90,8%, precisión 70,1%), con auditoría humana y segunda persona revisora (kappa 0,50). El mapa de 881 estudios, clasificados por título y resumen con instrumento validado contra 40 clasificaciones ciegas, muestra evaluación realizada por investigadores (67,4%), en entornos académicos y antes de la producción, la fase menos documentada (16,7%); métodos basados en datos de uso o ejecución aparecen en 7,7% y especialistas de dominio participan en 21,0%. La lectura íntegra de 46 estudios centrales organiza los desafíos en diez temas: la dificultad comienza con la definición de respuesta correcta y se repite en cada instrumento que intenta sustituirla. Ningún modelo individual, ni la dupla humana (kappa 0,50), alcanzó la concordancia de referencia para cribado por revisor único; el test-retest del autor dio kappa 0,57. Todos divergieron del autor en la misma dirección, y la calibración con ejemplos se asoció a mayor concordancia, en una muestra que sobrerrepresenta los casos límite.

PALABRAS CLAVE: *Evaluación de sistemas basados en LLM. Revisión de alcance. Ingeniería de software.*

1. INTRODUÇÃO

Sistemas de *software* cuja capacidade central é um modelo de linguagem de grande porte (LLM) tornaram-se uma categoria estabelecida de produto: assistentes de programação, *chatbots* de domínio, sistemas de geração aumentada por recuperação e agentes autônomos. Esses sistemas produzem respostas distintas para a mesma entrada, e a noção de resultado correto raramente admite especificação completa; o teste de *software* tradicional pressupõe comportamento repetível. Em entrevistas com 26 engenheiros que integravam funcionalidades baseadas em LLM a produtos comerciais, testar e avaliar o comportamento dos modelos figurou entre os maiores desafios cotidianos (PARNIN *et al.*, 2025).

Indústria e academia responderam com práticas novas de avaliação: conjuntos de referência (*golden datasets*), avaliação automatizada por LLM (*LLM-as-a-judge*), avaliação humana estruturada, testes A/B e monitoramento em produção. A literatura que as documenta é

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

recente (81% dos registros do *corpus* de triagem deste trabalho foram publicados em 2025 e 2026) e está dispersa entre engenharia de *software*, interação humano-computador e processamento de linguagem natural, comunidades que descrevem o mesmo fenômeno com vocabulários distintos (Seção 3.1).

A motivação parte também da experiência do primeiro autor no desenvolvimento de uma plataforma de análise baseada em LLM, na qual o conhecimento de domínio foi o gargalo da avaliação, percepção que coincide com achados recentes (MADEN *et al.*, 2026). Dela decorrem a hipótese de trabalho, em duas partes (o especialista de domínio participa da avaliação e, quando participa, é o gargalo), e as cinco perguntas que orientam a síntese dos desafios (Seção 4.7): como construir e manter conjuntos de referência; como rotular; como observar o comportamento nas junções entre os componentes do sistema (recuperação de documentos, *prompt*, modelo, *parsing* da saída e orquestração); quem pode julgar uma resposta de domínio; e o que as portas de verificação (*guardrails*, asserções, filtros e oráculos automáticos) deixam passar.

A pergunta de pesquisa é: como a avaliação de produtos de *software* baseados em LLM é praticada ao longo do ciclo de vida, segundo a literatura publicada a partir de 2022? O objetivo geral, mapear sistematicamente essa prática, desdobra-se em cinco objetivos específicos: identificar os métodos e métricas de avaliação reportados; caracterizar as fases do ciclo de vida em que cada prática se aplica; identificar os atributos de qualidade avaliados; sistematizar os desafios e lacunas reportados; e caracterizar quem executa a avaliação, com atenção ao papel de especialistas de domínio. Como resultado secundário, a triagem por instrumento baseado em LLM gera evidência sobre o desempenho desses modelos nessa tarefa (Seção 4.8). A cobertura inicia-se em 2022 porque a categoria de produto estudada nasce com a disponibilização em larga escala de modelos ajustados por instrução, consolidada com o lançamento do ChatGPT em novembro daquele ano.

2. REFERENCIAL TEÓRICO

2.1. Da avaliação de modelos à avaliação de produtos

A avaliação de LLMs consolidou-se primeiro no nível do modelo isolado, por meio de *benchmarks*, baterias padronizadas de tarefas com gabarito. Essa tradição não responde à pergunta de quem constrói produto: se o sistema que integra o modelo atende seus usuários com a qualidade esperada. No nível do sistema, a revisão mais próxima é a de Yu *et al.* (2025), que mapearam métricas e procedimentos para aferir a qualidade das saídas de sistemas de

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

inteligência artificial generativa, sem examinar em que fase do ciclo de vida a avaliação ocorre, quem a executa, nem o que acontece após a implantação.

2.2. A dimensão da prática

Shankar *et al.* (2024) demonstraram que mesmo a avaliação automatizada exige alinhamento iterativo com critérios humanos, formulando o problema de validar os validadores. Maden *et al.* (2026), em entrevistas com 19 profissionais, catalogaram dez práticas de avaliação em uso e identificaram um descompasso entre coletar resultados de avaliação e conseguir convertê-los em melhorias no produto. Essa dimensão, a prática como trabalho, segue sem sistematização.

2.3. Revisões adjacentes e posicionamento metodológico

Os estudos secundários existentes são parciais: além do mapeamento de métricas de saída (YU *et al.*, 2025), há revisões dedicadas às técnicas de LLM como juiz (LI *et al.*, 2024), à avaliação de agentes autônomos (YEHUDAI *et al.*, 2026) ou a um único domínio de aplicação, como a saúde. Nenhuma delas cruza as literaturas de engenharia de *software* e de interação humano-computador, cobre o ciclo de vida completo ou examina quem executa e como se organiza o trabalho de avaliação.

O desenho adotado pertence à tradição dos estudos de mapeamento sistemático da engenharia de *software* (PETERSEN *et al.*, 2008; PETERSEN; VAKKALANKA; KUZNIARZ, 2015), que classificam um campo emergente sem síntese agregada. O método de revisão de escopo remonta a Arksey e O'Malley (2005), e o relato segue a lista de verificação PRISMA-ScR (TRICCO *et al.*, 2018). A busca segue as diretrizes de revisões sistemáticas em engenharia de *software* (KITCHENHAM; CHARTERS, 2007) e o procedimento de *snowballing* (busca por citações para frente e para trás a partir de estudos conhecidos) de Wohlin (2014). A triagem com apoio de modelos de linguagem é prática recente: a literatura empírica mostra dependência do modelo e do *prompt* (KHRAISHA *et al.*, 2024; LÓPEZ-PINEDA *et al.*, 2025), e as organizações de síntese de evidências condicionam o uso a validação, supervisão humana e reporte integral (FLEMYNG *et al.*, 2025).

3. METODOLOGIA

Este trabalho é uma revisão de escopo relatada conforme a lista de verificação PRISMA-ScR (TRICCO *et al.*, 2018). O protocolo completo foi registrado na plataforma Open Science Framework (OSF) em 14 de agosto de 2026, antes do início de qualquer triagem, e está disponível em https://osf.io/2pmtw. A iteração própria das revisões de escopo (LEVAC;

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

COLQUHOUN; O'BRIEN, 2010) foi tratada por emendas datadas e por gatilhos fixados antes de cada auditoria. Foram registradas cinco emendas: a primeira passou a condução da triagem ao instrumento, com verificação do autor; a segunda acrescentou o segundo nível do instrumento; a terceira iniciou a retriagem manual; a quarta concluiu a triagem por regra instrumentada; e a quinta definiu a segunda pessoa revisora, o teste-reteste, a extração em duas camadas e a rodada limitada de *snowballing*. As três primeiras foram depositadas no OSF antes de sua aplicação; a quarta e a quinta foram aprovadas pelo segundo autor em 14 de outubro de 2026, antes de qualquer decisão delas decorrente, e depositadas no OSF em 17 de novembro de 2026.

3.1. Estratégia de busca

A pergunta de pesquisa foi estruturada no arcabouço PCC (população, conceito e contexto): população, produtos de *software* que integram LLMs; conceito, práticas, métodos e ferramentas de avaliação; contexto, literatura de engenharia de *software*, interação humano-computador e processamento de linguagem natural aplicado, publicada a partir de 2022. No piloto do protocolo, uma única expressão de busca no vocabulário de engenharia de *software* capturou apenas três dos dez estudos de referência, e nenhum dos sete escritos no vocabulário de interação humano-computador, de ferramentas e de prática. Por essa razão, a busca é híbrida, em três braços: (i) busca por títulos na base OpenAlex (PRIEM; PIWOWAR; ORR, 2022); (ii) enumeração das conferências de interação humano-computador CHI, UIST, IUI e CSCW na base DBLP; e (iii) *snowballing* a partir dos dez estudos de referência (WOHLIN, 2014). As três buscas foram executadas em 14 de agosto de 2026, com cobertura de publicações a partir de 1º de janeiro de 2022. No braço (i), a expressão, aplicada ao título, foi:

("LLM-based" OR "LLM-integrated" OR "LLM-powered" OR "LLM application") AND ("evaluation" OR "evals" OR "testing" OR "quality assurance" OR "monitoring")

No braço (ii), cada conferência foi consultada com as combinações de *evaluat*, *test*, *quality*, *monitor* e *benchmark* com *llm* e *language model*. No braço (iii), foram coletadas, no grafo de citações do OpenAlex, as referências publicadas a partir de 2022 e as citações de cada estudo de referência. A estratégia foi validada contra o conjunto de referência antes do registro do protocolo. A deduplicação foi feita por identificador e, em seguida, por título normalizado, para unir as versões *preprint* e publicada do mesmo estudo. Os registros recuperados, com data e proveniência por braço, estão arquivados no projeto do protocolo.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

3.2. Critérios de elegibilidade

São incluídos estudos publicados a partir de 2022, em inglês ou português, cujo objeto é a avaliação de um sistema ou produto de *software* que integra um LLM: estudos empíricos, propostas de técnica ou ferramenta de avaliação e relatos de experiência industrial, em veículo revisado por pares ou como *preprint* no arXiv ou em repositório equivalente (Zenodo, SSRN, Research Square, TechRxiv). São excluídos o *benchmarking* de modelos isolados sem contexto de aplicação, o uso de LLM para avaliar artefatos que não são produtos baseados em LLM, a avaliação de capacidades do modelo sem sistema de *software* ao redor, editoriais sem dados e a literatura cinzenta industrial, que pode ser citada na discussão, mas não compõe o *corpus*. Quatro regras de fronteira, refinadas na calibração do instrumento de triagem, tratam os casos difíceis: estudos cuja contribuição é uma avaliação e ferramentas de apoio ao fluxo de avaliação são incluídos; artigos de sistema cuja avaliação é autovalidação incidental da própria contribuição são excluídos; propostas de métrica validadas apenas em *benchmarks* de geração de linguagem, sem contexto de produto, são excluídas; e avaliações longitudinais de produtos implantados são incluídas. Estudos secundários que sintetizam práticas de avaliação são retidos no *corpus* (Seção 4.2).

3.3. Triagem

A triagem de título e resumo foi conduzida por um instrumento automatizado, que aplica os critérios da Seção 3.2 por meio de um modelo de linguagem e grava, para cada registro, uma decisão (inclusão, exclusão ou incerteza), o critério aplicado e uma justificativa curta. Dos 1.681 registros, 387 (23,0%) não tinham resumo nos metadados; 153 tiveram o resumo recuperado de outras fontes e foram retriados com ele, e os demais foram triados apenas pelo título. O instrumento registrado opera em dois níveis: o primeiro tria todos os registros; o segundo aplica um *prompt* mais restritivo às inclusões e às incertezas do primeiro, para separar a avaliação que é contribuição da avaliação incidental (Figura 1). Neste texto, “estrato de fronteira” designa os registros que passaram pelo segundo nível; “estrato de primeiro nível”, as exclusões do primeiro nível que o segundo não examinou; e “caso-limite”, o registro cuja elegibilidade depende das regras de fronteira. As decisões humanas foram registradas em uma interface de verificação que oculta, nas fases cegas, o veredito do instrumento.

A triagem combinou decisões do autor[3] e uma regra instrumentada, validada contra essas decisões e auditada por amostra. Na auditoria cega da primeira geração do instrumento (o instrumento registrado), sobre amostra de 230 registros, a taxa de reversão das exclusões foi

[3] Neste texto, “autor” refere-se ao primeiro autor, que tomou todas as decisões humanas de triagem e classificação; as demais inclusões são da regra validada e da triagem adicional da segunda pessoa revisora.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

de 45,8% no estrato de fronteira e de 11,3% no de primeiro nível; o resultado acionou o gatilho de retriagem manual completa, e o autor retriou 170 registros antes de a quarta emenda substituir a retriagem pela regra instrumentada. Os 230 rótulos da auditoria, os 23 da calibração da primeira geração e os 170 da retriagem formam as 423 decisões humanas (25,2% do *corpus* de triagem) que calibraram e validaram o instrumento de segunda geração.

O instrumento de segunda geração acrescenta 30 exemplos extraídos das decisões do autor. Foi validado, antes do uso, sobre os 393 registros decididos pelo autor que não serviram de exemplo, com três modelos do mesmo fornecedor e dois de outros fornecedores sob o mesmo *prompt* (Material suplementar S1). Nenhum modelo isolado alcançou *kappa* de Cohen de 0,8, referência citada pela Cochrane para autorizar triagem por uma só pessoa (GARRITTY *et al.*, 2024), e todos divergiram do autor na mesma direção: excluíram casos-limite em que a avaliação é a própria contribuição do estudo (Seção 4.8). A regra de inclusão foi fixada a partir dessa validação: um registro é incluído se pelo menos um de três modelos (claude-sonnet-5, claude-opus-5 e deepseek-v4-pro) o inclui; se nenhum inclui, o registro é excluído, mesmo que algum modelo o classifique como incerteza. A regra privilegia a revocação (*recall*) porque uma exclusão indevida na triagem não é recuperada depois: a leitura em texto integral restringe-se à camada profunda (Seção 3.4). A regra foi aplicada aos 1.258 registros sem decisão do autor e, em uma rodada limitada de *snowballing*, às citações diretas dos estudos da camada profunda; as inclusões dessa rodada não têm auditoria própria.

A reprodutibilidade das decisões foi medida por uma segunda pessoa revisora, calibrada em dez registros e sem acesso às decisões do autor. Ela triou às cegas 85 registros: 45 sorteados entre os 423 decididos pelo autor e 40 sorteados entre as exclusões da regra. O desenho adapta a recomendação da Cochrane para revisões rápidas (GARRITTY *et al.*, 2024): a medida é posterior à triagem, cobre 5% do *corpus* de triagem e adota o limiar de *kappa* de 0,6, reportado com intervalo de confiança (IC). O limiar é menor que o exigido dos modelos porque a medida verifica se as decisões do autor são reprodutíveis por outra pessoa. A quinta emenda fixou a consequência de antemão: *kappa* abaixo de 0,6 nos 45 registros decididos pelo autor estende a segunda triagem às exclusões do estrato de fronteira, sem alteração retroativa de rótulos. Como controle adicional, 70 exclusões da regra receberam rótulo humano (as 40 da amostra da segunda pessoa revisora e 30 rotuladas pelo autor, misturadas a 10 iscas, registros sabidamente elegíveis), e a falsa inclusão foi medida nos 40 estudos da validação cega da classificação do mapa (Seção 3.4). Os gatilhos foram fixados antes das auditorias: mais de 8 falsas exclusões em 70 tornariam a regra mais inclusiva; mais de 40% de falsas inclusões fariam o mapa ser

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

reportado também sob a regra de dois votos, em que um registro só entra se ao menos dois dos três modelos o incluem.

3.4. Extração e síntese

A extração ocorreu em duas camadas, opção que o protocolo previa, por emenda registrada, para mais de cem estudos incluídos. Na camada larga, todos os estudos incluídos foram classificados a partir do título e do resumo quanto a tipo de contribuição, domínio de aplicação, tipo de sistema, métodos de avaliação, fase do ciclo de vida, atributos de qualidade, quem avalia, métricas nomeadas e cenário, com categorias fixadas no Apêndice D do protocolo. As fases são três: desenvolvimento, quando a avaliação orienta a iteração sobre *prompts*, modelos e componentes; pré-implantação, quando ela sustenta a decisão de liberar o produto; e produção, quando o produto já está em uso real. Contam-se como *online* os métodos que analisam o comportamento do produto com dados de uso ou de execução (testes A/B, monitoramento em produção, análise de registros de uso e mineração de *feedback*), incluindo sessões experimentais. A classificação foi pré-preenchida por um instrumento de classificação (o claude-opus-5) e validada contra 40 estudos classificados às cegas pelo autor em cinco campos (tipo de contribuição, cenário, métodos de avaliação, fase do ciclo de vida e quem avalia), além da elegibilidade. O autor releu os resumos dos registros com fase ou cenário indeterminados e reclassificou 36 fases e 128 cenários; nos 40 estudos da validação e nos 46 da camada profunda, consultou também a introdução e a conclusão.

Na camada profunda, os estudos centrais à pergunta de pesquisa (os que examinam a própria prática de avaliação e os que propõem ferramentas ou técnicas de apoio à avaliação) foram lidos integralmente pelo autor e extraídos por formulário estruturado com 19 campos alinhados aos objetivos específicos, entre eles métodos de avaliação, fase do ciclo de vida, atributos de qualidade, quem avalia, envolvimento de especialistas de domínio e desafios reportados (com citação literal); os 19 campos e seus valores admitidos estão no Apêndice E do protocolo. A seleção seguiu regra declarada: os estudos de prática, até o teto de 20, sorteados por ano quando havia mais e verificados em texto integral; e uma amostra sorteada de estudos de ferramentas, estratificada por ano e por situação do veículo e ampliada na ordem do sorteio até 27, quando os temas se repetiram sem acrescentar categoria nova. Quatro estudos sorteados foram substituídos por falta de acesso ao texto integral, pela próxima entrada da fila do sorteio.

A extração foi pré-preenchida pelo instrumento de extração com a citação literal que sustenta cada campo, e as 46 fichas foram verificadas pelo autor: três extraídas às cegas, nove conferidas valor a valor (127 dos 135 campos de conteúdo confirmados, 8 corrigidos) e 34 por checagem de consistência interna (150 dos 204 campos usados na síntese confirmados, 48

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

marcados como dúvida e 6 como correção; as dúvidas foram revisadas e as correções aplicadas antes da síntese). Na síntese temática, o autor agrupou as citações literais de desafios que tratavam da mesma dificuldade e nomeou cada grupo como tema. A elegibilidade em texto integral foi avaliada na camada profunda; para os estudos que entram apenas no mapa, foi estimada pela validação cega da classificação (Seção 4.1). A síntese é descritiva e temática; conforme o padrão de revisões de escopo, não há meta-análise nem avaliação formal de risco de viés dos estudos primários (TRICCO *et al.*, 2018).

3.5. Uso de modelos de linguagem no método

Os julgamentos de triagem e classificação dos cinco campos validados (Seção 4.2) sugeridos por modelo de linguagem seguem as condições da posição conjunta das organizações de síntese de evidências (FLEMYNG *et al.*, 2025): responsabilidade integral do autor, validação antes do uso, supervisão humana, reporte de todo julgamento feito por instrumento e publicação de modelo, versão, *prompts* e vereditos. Os modelos são referidos por identificadores fixos: claude-sonnet-5 e claude-opus-5 foram invocados pela interface de linha de comando Claude Code, versão 2.1.236, em lotes de dez registros; o claude-fable-5-1, pela mesma interface por mecanismo distinto, em lotes de 15; deepseek-v4-pro e gpt-5.4, por API, com temperatura zero. Os *prompts* integrais (Apêndices A a E do protocolo) e os vereditos de cada execução, registro a registro, estão arquivados no projeto do protocolo; a extração da camada profunda foi pré-preenchida pelo claude-opus-5. Só os três modelos da regra (Seção 3.3) têm poder de decisão. As fontes de financiamento dos estudos incluídos não foram extraídas.

4. RESULTADOS

4.1. Fluxo da seleção

A busca em três braços identificou 2.116 registros únicos por identificador (1.343 pela expressão de busca, 46 pela enumeração das conferências e 747 pelo *snowballing*; os braços se sobrepõem), reduzidos a 1.681 após a deduplicação por título. O autor decidiu 423 registros (164 inclusões, 229 exclusões e 30 incertezas; as incertezas não entram no *corpus*); a regra validada incluiu 506 dos registros restantes (110 do estrato de primeiro nível e 396 da fronteira) e excluiu 752. A rodada limitada de *snowballing* identificou 737 candidatos fora do *corpus*, dos quais a mesma regra incluiu 205. O *corpus* do mapa tem 881 estudos: 164 decididos pelo autor, 506 pela regra na busca principal, 205 pela regra no *snowballing* e 6 identificados na triagem adicional das exclusões de fronteira (Figura 1). Os motivos das 229 exclusões do autor foram: autovalidação incidental da própria contribuição, 88; *benchmarking* de modelo sem contexto de aplicação, 35; editorial ou artigo de posição sem dados, 34; capacidade do modelo sem sistema

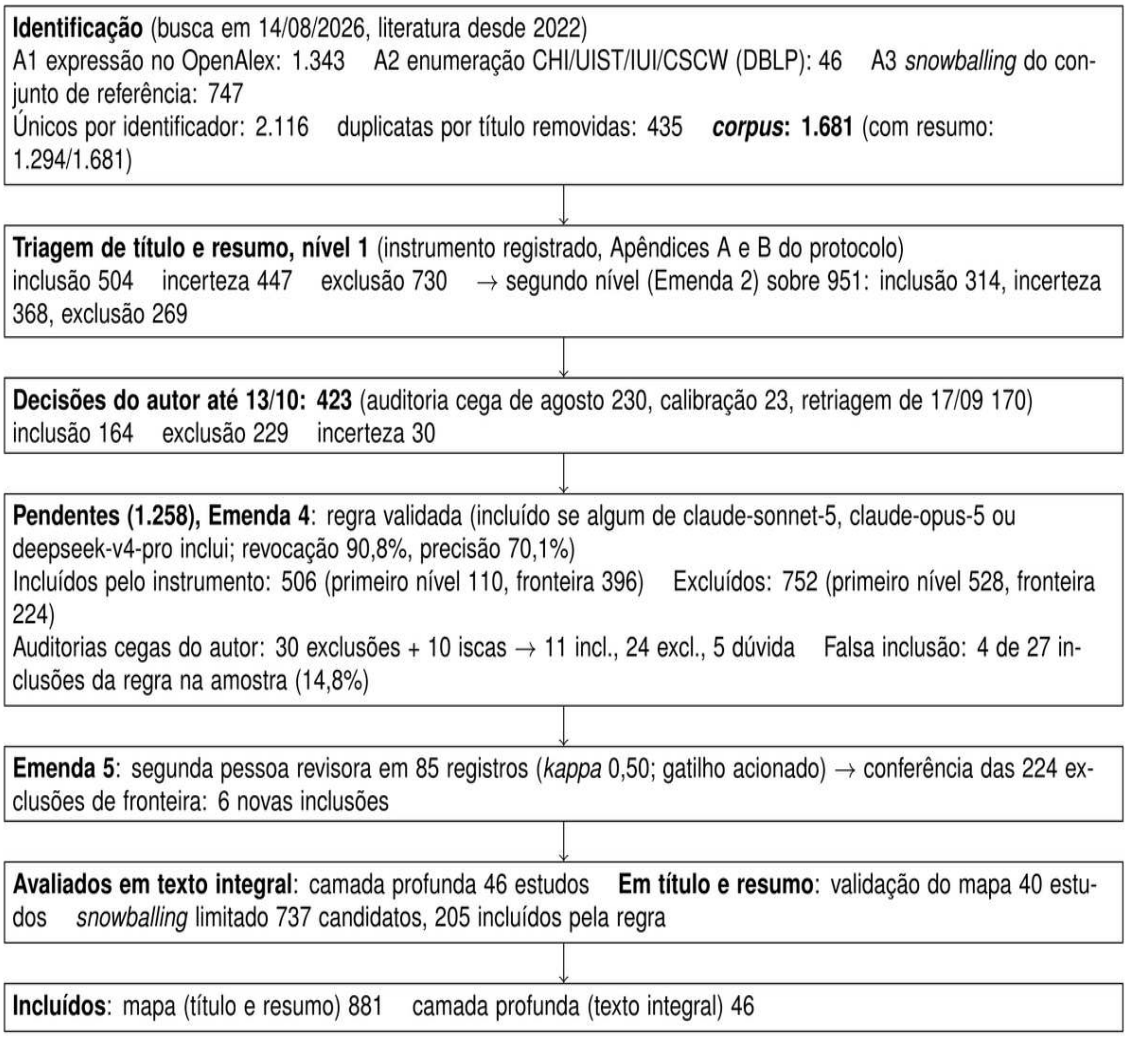
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

ao redor, 31; LLM avaliando artefato que não é produto baseado em LLM, 24; métrica sem contexto de produto, 14; e 3 sem critério registrado. As exclusões da regra não têm motivo individual: são os registros que nenhum dos três modelos incluiu. Na camada profunda, os 46 estudos lidos dividem-se em 19 de prática e 27 de ferramentas ou técnicas.

A segunda pessoa revisora concordou com o autor em 29 dos 45 registros nas três classes (64,4%) e em 34 dos 45 na decisão binária incluir/não incluir (75,6%; *kappa* de Cohen de 0,50; IC de 95%, por *bootstrap*, de 0,25 a 0,73), abaixo do limiar de 0,6. O gatilho correspondente foi acionado: a revisora triou as 224 exclusões de fronteira, com apoio do instrumento de triagem para apresentar os registros, e identificou 6 como inclusões, incorporadas ao *corpus* sem alteração retroativa dos rótulos originais. Em 6 das 16 discordâncias nas três classes, o autor incluiu como avaliação que é contribuição o que a revisora excluiu como autovalidação incidental; as demais envolvem a classe de incerteza. Nas 70 exclusões da regra com rótulo humano, a revisora reverteu 1 e o autor reverteu 1 e marcou 5 como incerteza; as 10 iscas foram reconhecidas como elegíveis. Dos 40 estudos do mapa classificados às cegas pelo autor, 27 foram confirmados elegíveis, 5 não elegíveis e 8 marcados como dúvida; entre as inclusões da regra na mesma amostra (também 27, por coincidência), 4 foram julgadas não elegíveis pelo autor (falsa inclusão de 14,8%, com as dúvidas contadas como elegíveis). O gatilho de falsa exclusão (Seção 3.3) não foi acionado, e o de falsa inclusão não foi acionado sob a convenção adotada (dúvida como elegível; o pior caso está na Seção 5.5).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figura 1. Fluxo da seleção dos estudos



Fonte: Elaborado pelos autores (2026) a partir dos registros arquivados da triagem.
Nota: Segue o item 14 da lista PRISMA-ScR, com as etapas não padrão desta revisão: vereditos do instrumento, decisões do autor por lote e por fase, exclusões exclusivas do instrumento com auditoria e avaliação em texto integral na camada profunda.

4.2. Caracterização do *corpus*

A classificação da camada larga foi validada em 40 estudos classificados às cegas pelo autor; os *kappas* a seguir são calculados sobre os 27 confirmados elegíveis. Nos campos de rótulo único, o *kappa* de Cohen foi de 0,62 para tipo de contribuição e de 0,63 para cenário. Nos campos de rótulo múltiplo, após alinhamento *post hoc* da granularidade (as subcategorias de "outro" geradas pelo instrumento foram colapsadas no rótulo-pai após a medição inicial), o *kappa* médio por rótulo de métodos de avaliação passou de 0,20 para 0,69, e o de quem avalia foi de

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

0,65. Para a fase do ciclo de vida, o *kappa* por rótulo na classificação fina, em três fases, foi de 0,41, insuficiente para distingui-las; a classificação binária (antes de produção *versus* produção), que sustenta a conclusão de que a produção é sub-representada, teve *kappa* de 0,66. As distribuições das seções seguintes são estimativas do instrumento de classificação, com exceção de 36 fases e 128 cenários reclassificados pelo autor (Seção 3.4).

Quanto ao ano de publicação e à situação do veículo (Figura S1 do material suplementar), 71,6% dos estudos foram publicados nos dois últimos anos, e 2026 cobre apenas até a data da busca; 663 estudos estão em periódicos ou anais, 86 apenas no arXiv e 53 em repositórios de *preprints*; o restante não teve o veículo determinado pelos metadados. Os domínios de aplicação mais frequentes são educação (189 estudos) e saúde (114), seguidos de assistente de programação (60).

Pelo tipo de contribuição, as categorias mais frequentes são propostas de ferramentas ou técnicas de avaliação (243; 27,6%), sistemas ou técnicas cuja avaliação é parte da contribuição (181; 20,5%), avaliações de produtos específicos (157; 17,8%), estudos com usuários (147; 16,7%), *benchmarks* aplicados a produtos (68; 7,7%) e estudos da prática de avaliação (46; 5,2%); entre os 39 restantes estão os 10 estudos secundários que sintetizam práticas de avaliação. O cenário é acadêmico em 617 estudos (70,0%), industrial em 94 (10,7%) e misto em 80 (9,1%); em 90 (10,2%), o resumo não permite determiná-lo.

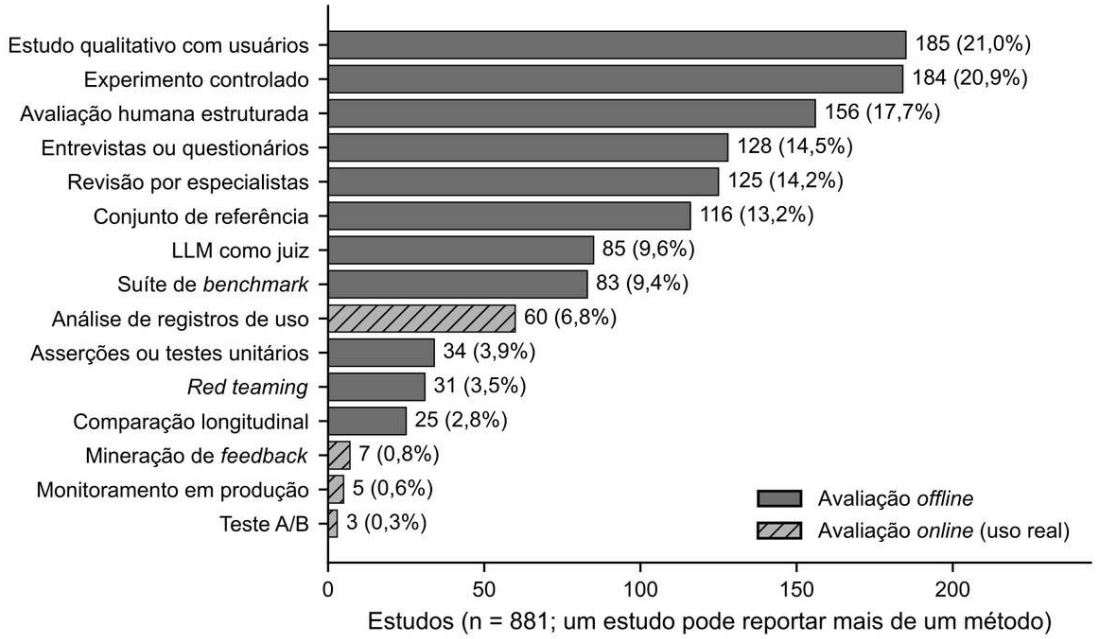
4.3. Métodos e métricas de avaliação

A Figura 2 apresenta a frequência de cada método. Os quatro mais frequentes vêm da pesquisa com usuários e do experimento: estudo qualitativo com usuários (185; 21,0%), experimento controlado (184; 20,9%), avaliação humana estruturada (156; 17,7%) e entrevistas ou questionários (128; 14,5%). Seguem a revisão por especialistas (125; 14,2%), o conjunto de referência (116; 13,2%), o LLM como juiz (85; 9,6%) e a suíte de *benchmark* (83; 9,4%). Métodos baseados em dados de uso ou de execução e métodos automatizados de engenharia (asserções ou testes unitários e *red teaming*, os ataques deliberados para provocar falhas) são raros: só a análise de registros de uso (60; 6,8%) passa de 5%, e apenas 68 estudos (7,7%) reportam ao menos um método *online* (Seção 3.4).

O LLM como juiz e a suíte de *benchmark* são os métodos com maior crescimento em contagem absoluta: o juiz passa de 0 estudos em 2022 a 42 em 2026, e o *benchmark*, de 11 em 2024 a 41 em 2026. As métricas nomeadas nos resumos somam 1.327 rótulos distintos após normalização simples, com forte dispersão terminológica; a acurácia é a mais citada (49 estudos).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figura 2. Métodos de avaliação reportados



Fonte: Elaborado pelos autores (2026) a partir do mapa da camada larga.
Nota: 881 estudos; um estudo pode reportar mais de um método.

4.4. Fases do ciclo de vida

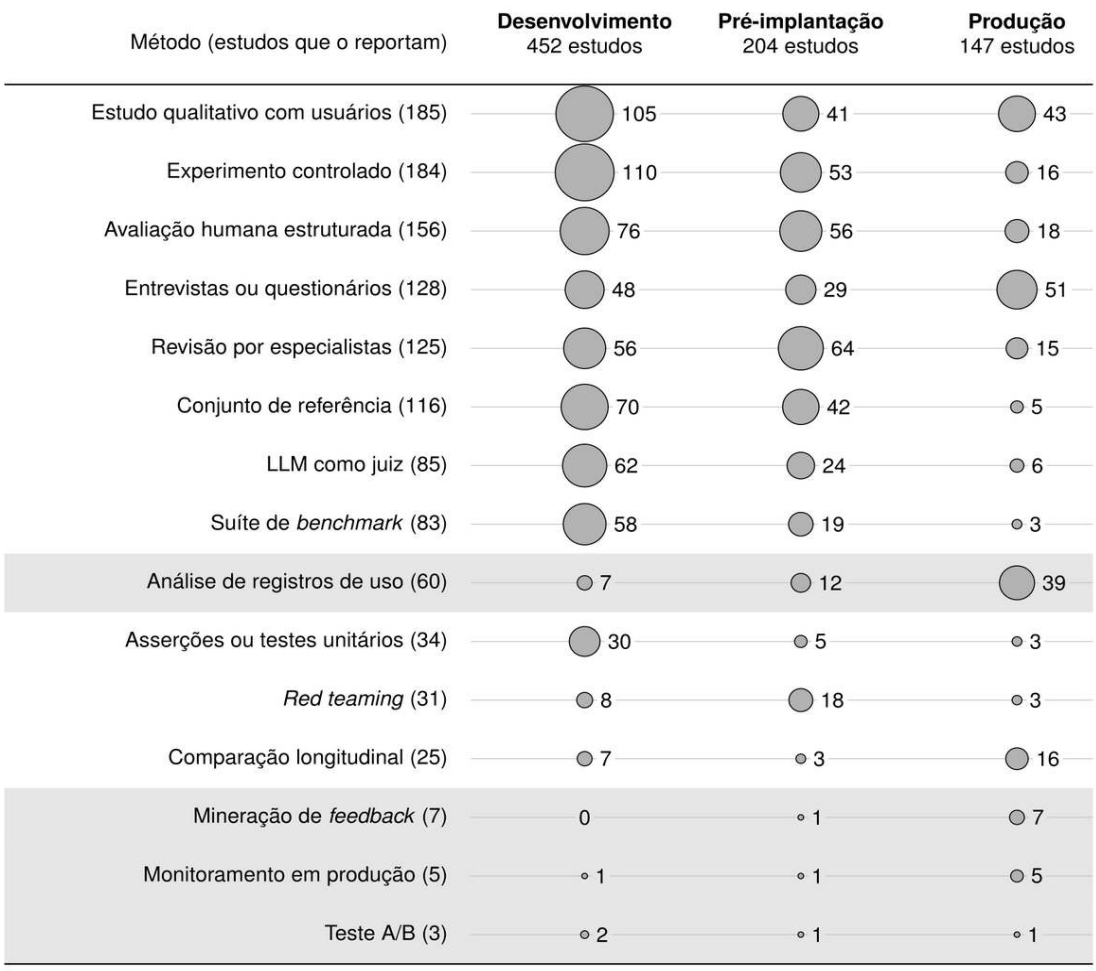
A fase de desenvolvimento concentra 452 estudos (51,3%), a pré-implantação, 204 (23,2%), e a produção, 147 (16,7%); em 147 estudos (16,7%), a fase não é determinável pelo resumo, e um estudo pode ser classificado em mais de uma fase. A produção é a menos documentada das três fases, conclusão que se apoia na classificação binária (*kappa* de 0,66; Seção 4.2); as contagens por fase que seguem usam a classificação fina (*kappa* de 0,41) e devem ser lidas como exploratórias.

A Figura 3 cruza os métodos com as fases, contando cada estudo em todos os pares que ele reporta; as mesmas contagens estão na Tabela S2 do material suplementar. Conjunto de referência (70 de 116 estudos), LLM como juiz (62 de 85), suíte de *benchmark* (58 de 83) e asserções (30 de 34) concentram-se no desenvolvimento. A revisão por especialistas (64 de 125) e a avaliação humana estruturada (56 de 156) são os métodos mais presentes na pré-implantação. Na produção, os métodos mais frequentes são as entrevistas ou questionários (51 dos 147 estudos na fase), o estudo qualitativo com usuários (43) e a análise de registros de uso (39), enquanto o conjunto de referência (5) e o LLM como juiz (6) quase desaparecem. A

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

literatura descreve a avaliação em produção como observação de uso e escuta de usuários, e raramente como medição contínua da qualidade das respostas.

Figura 3. Métodos de avaliação por fase do ciclo de vida



Fonte: Elaborado pelos autores (2026) a partir do mapa da camada larga.
Nota: 881 estudos; a área de cada círculo é proporcional ao número de estudos que reportam o método na fase; um estudo conta em cada par método–fase que reporta; as faixas cinza marcam os métodos que observam o uso real.

4.5. Atributos de qualidade

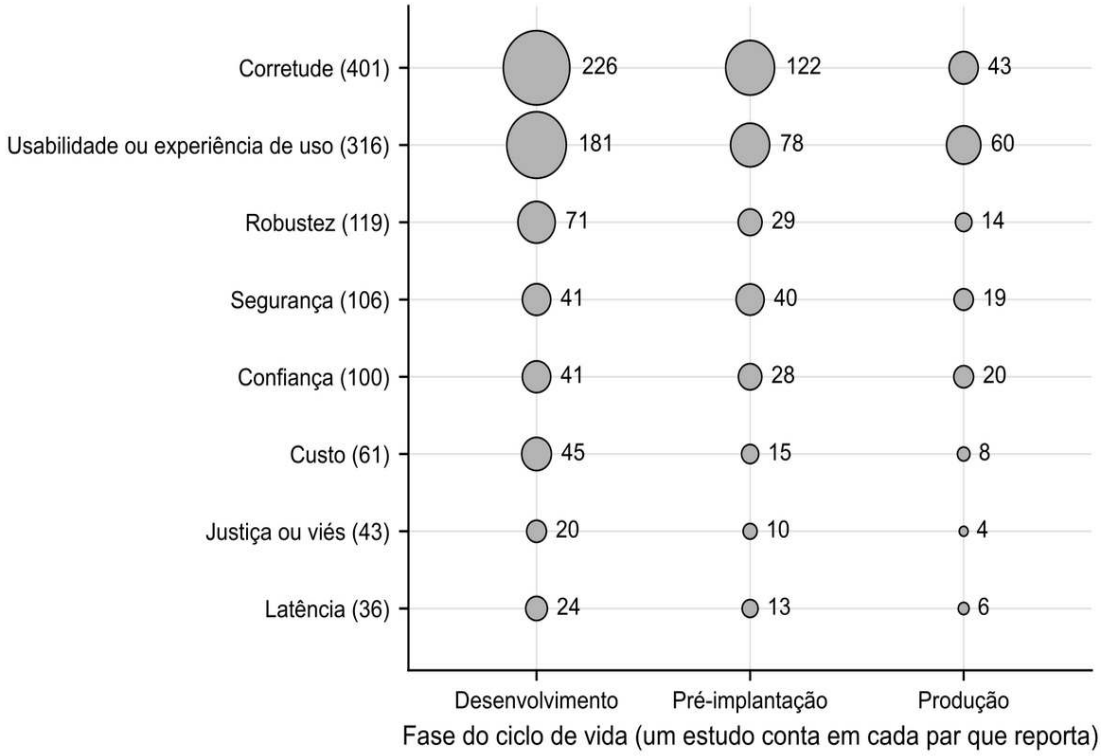
Corretude (401; 45,5%) e usabilidade ou experiência de uso (316; 35,9%) dominam os atributos avaliados; seguem robustez (119; 13,5%), segurança (106; 12,0%), confiança (100; 11,4%), custo (61; 6,9%), justiça ou viés (43; 4,9%) e latência (36; 4,1%). A Figura 4 cruza os atributos com as fases.

A corretude concentra-se no desenvolvimento (226 dos 401 estudos) e cai na produção (43). Entre os oito atributos principais, a segurança é o único com número praticamente igual de

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

estudos na pré-implantação e no desenvolvimento (40 e 41). A usabilidade é o atributo mais presente na produção (60 estudos). Custo e latência, atributos operacionais, aparecem em menos de um estudo em dez e são medidos sobretudo no desenvolvimento (45 de 61 e 24 de 36), raramente em produção (8 e 6). O campo não tem validação por amostra: as contagens são estimativas do instrumento de classificação.

Figura 4. Atributos de qualidade por fase do ciclo de vida



Fonte: Elaborado pelos autores (2026) a partir do mapa da camada larga.
Nota: 881 estudos; a área de cada círculo é proporcional ao número de estudos que avaliam o atributo na fase.

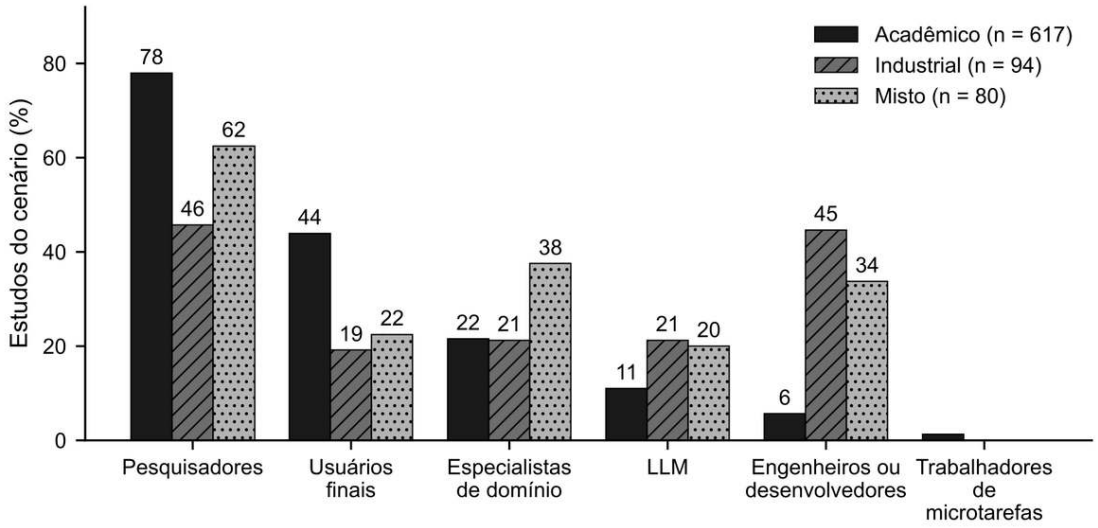
4.6. Quem avalia e o papel dos especialistas de domínio

Pesquisadores conduzem a avaliação em 594 estudos (67,4%), usuários finais participam em 311 (35,3%), especialistas de domínio em 185 (21,0%), um LLM avalia em 110 (12,5%) e engenheiros ou desenvolvedores em 108 (12,3%). Quem avalia muda com o cenário (Figura 5): entre os 94 estudos em cenário industrial, engenheiros avaliam em 42 e pesquisadores em 43; no cenário acadêmico, pesquisadores conduzem a avaliação em quase quatro de cada cinco estudos (481 de 617) e os usuários finais participam em 43,9% deles, contra

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

19,1% na indústria. A primeira parte da hipótese da Seção 1 encontra apoio localizado: o especialista de domínio está presente em um quinto dos estudos, com concentração na saúde (67 dos 114 estudos do domínio).

Figura 5. Quem conduz a avaliação, por cenário do estudo



Fonte: Elaborado pelos autores (2026) a partir do mapa da camada larga.
Nota: Percentual dos estudos de cada cenário; os 90 estudos de cenário indeterminado não entram na figura.

Dos 46 estudos lidos integralmente, 21 (45,7%) não envolvem nenhum especialista do domínio de aplicação na avaliação reportada. Nos 25 restantes, os papéis acumulam-se: o especialista julga, pontua ou adjudica saídas em 16 estudos; constrói ou rotula o conjunto de referência em 7; define ou revisa critérios, rubricas e taxonomias em 7; e participa só de forma formativa em 7.

4.7. Desafios reportados

O quarto objetivo é respondido pela camada profunda, porque o resumo raramente nomeia o que não funcionou. A leitura das 46 fichas de extração, por codificador único (o autor), partiu das cinco perguntas da Seção 1 e produziu dez temas: cinco correspondem às perguntas (oráculo ausente, conjuntos de referência, especialista, portas de verificação e junções do *pipeline*, isto é, da cadeia de componentes que processa cada requisição) e cinco emergiram de desafios não cobertos por elas (método não consolidado, escala, não determinismo, LLM como juiz e sinal de produção). O Quadro 1 resume os temas, em ordem de frequência, com exemplos; a Figura S2 do material suplementar os dispõe em cinco grupos, com as relações entre eles.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Os temas de método e oráculo antecedem qualquer medição. Treze dos 19 profissionais entrevistados por Maden *et al.* (2026, p. 10) não conseguem definir o que é bom antes de liberar o produto. Quando os critérios existem, eles mudam durante a avaliação (GEBREEGZIABHER *et al.*, 2025); não há métrica de adequação de testes aceita (PAN *et al.*, 2026), nem limiar de bom o suficiente (PARNIN *et al.*, 2025; NAVARRO; SYRIANI; ARAWJO, 2026).

Quadro 1. Temas de desafios reportados na camada profunda

Tema	***n***	**Exemplos**
Método não consolidado: sem métrica de adequação nem limiar de bom o suficiente; cada equipe inventa o seu; propostas acadêmicas não testadas em produto	26	(MADEN *et al.*, 2026; PAN *et al.*, 2026; PARNIN *et al.*, 2025; JIANG *et al.*, 2022; NAVARRO; SYRIANI; ARAWJO, 2026; VANEL *et al.*, 2024)
Oráculo de teste ausente: não há mecanismo que diga se uma saída está correta, o que é bom é subjetivo e os critérios mudam durante a avaliação	25	(GEBREEGZIABHER *et al.*, 2025; MADEN *et al.*, 2026; NAHAR *et al.*, 2025; GARCIA *et al.*, 2026; SAH *et al.*, 2026; ZAMFIRESCU-PEREIRA *et al.*, 2023)
Conjuntos de referência caros, pequenos ou perecíveis; *benchmarks* públicos irrelevantes ao caso de uso ou contaminados (com itens que entraram nos dados de treinamento do modelo); dados sintéticos como substituto	25	(PARNIN *et al.*, 2025; QIAN *et al.*, 2025; SORDO *et al.*, 2026; GUO *et al.*, 2026; VERO *et al.*, 2026; ZHAO *et al.*, 2025)
Avaliação manual que não escala: custo, tempo, carga mental e fadiga de quem revisa	24	(ALBUQUERQUE *et al.*, 2024; PARNIN *et al.*, 2025; WANG *et al.*, 2025; COSCIA; ENDERT, 2024; HE *et al.*, 2025)
Não determinismo, sensibilidade ao *prompt* e deriva do modelo, que quebram a reprodutibilidade e os testes	22	(GAJJAR, 2025; MAGALHÃES *et al.*, 2025; POTTERS; HOFMAN, 2026; LIU *et al.*, 2024)
Junções do *pipeline*: integração, *parsing*, falhas em cascata, caixa-preta, métricas agregadas que mascaram falhas, resultados não acionáveis	20	(MAGALHÃES *et al.*, 2025; MULIAN *et al.*, 2026; TIAN *et al.*, 2026)
LLM como juiz com viés, inconsistência ou pontos cegos correlatos ao sistema avaliado, sem validação	19	(LIU *et al.*, 2025; SZYMANSKI *et al.*, 2026; TAN *et al.*, 2024)
Portas de verificação (*guardrails*, asserções, filtros, oráculos automáticos) que deixam passar erros ou bloqueiam demais	17	(GAJJAR, 2025; GOH *et al.*, 2025; NAHAR *et al.*, 2025)
Dependência do especialista de domínio, que é caro, escasso, inconsistente ou não articula seus critérios	15	(DEMYANENKO, 2026; QIAN *et al.*, 2025; SZYMANSKI *et al.*, 2026; LAURO; BIGHAM; KOTTURI, 2024)

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Tema	*n*	Exemplos
Sinal de produção fraco ou inacessível: *feedback* binário, disparidade *offline*/*online*, detecção lenta, registros restritos	11	(MADEN *et al.*, 2026; MAILACH *et al.*, 2025; NAHAR *et al.*, 2025)

Fonte: Elaborado pelos autores (2026) a partir das fichas de extração da camada profunda.
Nota: 46 estudos; um estudo pode sustentar vários temas.

Os conjuntos de referência são descritos como caros de construir e perecíveis: os exemplos curados precisam ser refeitos a cada mudança de *prompt* ou de modelo (PARNIN *et al.*, 2025; VERO *et al.*, 2026). A rotulagem depende do especialista de domínio em 15 estudos: consultores pedagógicos abandonaram a pontuação rotineira por tédio (SZYMANSKI *et al.*, 2026), oftalmologistas discordaram entre si ao ranquear respostas (TAN *et al.*, 2024) e especialistas de longa carreira reconhecem o conteúdo bom sem conseguir articular os critérios (DEMYANENKO, 2026).

Com o sistema em exercício, o não determinismo faz cada caso de teste se comportar como um teste intermitente (MAGALHÃES *et al.*, 2025; PARNIN *et al.*, 2025; HE *et al.*, 2025), e pequenos ajustes no *prompt* ou no modelo invalidam os testes de regressão (NAHAR *et al.*, 2025; JIANG *et al.*, 2022). O LLM como juiz, resposta recorrente ao custo da avaliação humana, aparece em 19 estudos como um problema em si. Concordou com clínicos no ranqueamento de respostas (correlação de Spearman de 0,90) e ainda assim não identificou uma resposta clinicamente perigosa que os clínicos apontaram (TAN *et al.*, 2024). Prefere saídas de modelos da mesma família que a sua, o chamado viés de linhagem (SORDO *et al.*, 2026; TAN *et al.*, 2024; YU, 2026); pode compartilhar pontos cegos com o sistema avaliado (GAJJAR, 2025; YU, 2026; GUO *et al.*, 2026); e é vulnerável a ataques que o levam a aprovar respostas erradas, com taxa de sucesso de até 99,8% no cenário mais desfavorável testado (LIU *et al.*, 2025).

Nas portas de verificação, os *guardrails* podem dar falsa sensação de segurança e, ao mesmo tempo, bloqueiam interações inócuas (NAHAR *et al.*, 2025), e nenhuma garantia teórica é possível dado o espaço de entradas (GOH *et al.*, 2025). O oráculo híbrido que substituiu as asserções de correspondência exata, em uma única aplicação móvel, reduziu o escape de defeitos de 14,2% para 6,8% sem eliminá-lo (GAJJAR, 2025). Nas junções do *pipeline*, falhas de *parsing* respondem por cerca de 38% das falhas de tarefa em sistemas baseados em agentes em produção (MULIAN *et al.*, 2026), e 17 dos 19 profissionais entrevistados não conseguem converter resultados em ações (MADEN *et al.*, 2026, p. 11). Fecham o quadro a avaliação manual que não escala (24 estudos) e o sinal de produção fraco (11 estudos).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

4.8. O instrumento de triagem como resultado

Sobre os 393 registros de validação, com 152 inclusões do autor, nenhuma das seis configurações testadas (cinco modelos de três fornecedores, um deles nas duas gerações do instrumento; Tabela S1, no Material suplementar S1) atingiu o *kappa* de 0,8 que a Cochrane cita como referência para autorizar triagem por uma só pessoa em revisões rápidas (GARRITTY *et al.*, 2024). Na decisão binária incluir/não incluir, com incerteza contada como não inclusão, o *kappa* ficou entre 0,21 (primeira geração) e 0,69 nos modelos de segunda geração (0,54 a 0,69). Para comparação, a concordância binária entre o autor e a segunda pessoa revisora (Seção 4.1) deu *kappa* de 0,50 (IC de 95%: 0,25 a 0,73); o intervalo contém as estimativas pontuais de todos os modelos de segunda geração, de modo que a diferença entre concordância humano-humano e humano-modelo não é distinguível nessa amostra. No teste-reteste previsto na quinta emenda, o autor decidiu de novo, às cegas, 40 registros sorteados entre os que já tinha decidido, com concordância incluir/não incluir de 80,0% (32 de 40) e *kappa* de 0,57 (IC de 95%: 0,29 a 0,82), o que limita o *kappa* alcançável contra este padrão de referência. Todos os modelos divergiram do autor na mesma direção, excluindo casos-limite em que a avaliação é a contribuição, e a fração de inclusões do autor que cada modelo excluiria sozinho variou de 19,1% a 34,9% entre os modelos de segunda geração. O conjunto de validação sobrerrepresenta os casos-limite, porque vem de lotes concentrados em exclusões de fronteira (Material suplementar S1).

A regra de inclusão adotada combina três modelos de dois fornecedores, com deepseek-v4-pro incluído por recuperar inclusões que claude-sonnet-5 e claude-opus-5 perdem. No conjunto de validação, a regra recupera 90,8% das inclusões do autor com precisão de 70,1%; entre as exclusões da regra, 7,1% são inclusões do autor, e entre as inclusões do autor, 9,2% são perdidas pela regra (14 de 152). A falsa inclusão é de 29,9% das inclusões da regra. Esses números dão a margem de leitura do mapa: entre os 711 registros incluídos por regra (80,7% do mapa), a fração que o autor excluiria foi estimada em 14,8% na amostra de validação do mapa (4 em 27; IC de 95%: 4% a 34%; Seção 4.1). Fora da validação, em 70 exclusões da regra auditadas às cegas, 2 foram revertidas (2,9%; IC de 95%: 0,3% a 9,9%).

5. DISCUSSÃO

Na literatura, a avaliação de produtos baseados em LLM é praticada sobretudo por pesquisadores, em cenário acadêmico e antes da produção, com métodos herdados da pesquisa com usuários e do experimento controlado. As distribuições são estimativas com erro medido para cinco dos nove campos (Seção 4.2), e a discussão apoia-se nas diferenças de maior magnitude. As práticas que profissionais descrevem como desenvolvimento orientado por

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

avaliação, com conjuntos de referência, juízes automáticos e monitoramento contínuo (MADEN *et al.*, 2026; NAHAR *et al.*, 2025), encontram base empírica estreita no *corpus*: os métodos baseados em dados de uso ou de execução somam 7,7% dos estudos.

5.1. Uma leitura pelo ciclo de vida

No início do ciclo, a dificuldade é saber o que medir. Dois dos três temas mais frequentes da camada profunda, oráculo ausente e método não consolidado, descrevem equipes que começam sem especificação do que conta como resposta correta e só a descobrem examinando saídas. A deriva de critérios observada por Shankar *et al.* (2024) e reencontrada aqui em contextos distintos sugere que a especificação completa do que é bom, pré-requisito do teste tradicional, não está disponível no início: ela resulta da avaliação. O manual para praticantes de Shankar e Husain, fora do *corpus* por ser literatura de praticantes (Seção 3.2), parte da mesma premissa (SHANKAR; HUSAIN, 2026, cap. 1 e 4). A convergência não é independente, porque Shankar é a primeira autora do estudo sobre deriva de critérios; o que esta revisão acrescenta é a evidência de que a condição se repete em estudos de outros grupos e domínios.

Antes da entrega, o conjunto de referência, instrumento que a camada profunda mais discute, perde valor quando o *prompt* ou o modelo muda (PARNIN *et al.*, 2025), e a pessoa capaz de rotular com autoridade é cara de recrutar (MADEN *et al.*, 2026) e se esgota antes de cobrir o volume (SZYMANSKI *et al.*, 2026). As portas de verificação reduzem o escape sem zerá-lo: o estudo de Gajjar (2025) conclui que o problema do oráculo foi convertido em um exercício de calibração, e nos estudos lidos nenhuma porta zerou o escape (GAJJAR, 2025; GOH *et al.*, 2025).

O ciclo deveria se fechar na produção, a fase que o mapa menos documenta (16,7% dos estudos). A camada profunda indica por quê: o sinal disponível é fraco, e as métricas *offline* e *online* divergem. Passada a porta, a dificuldade seguinte é a atribuição: saber qual componente do *pipeline* produziu a falha exige observar as junções entre os componentes, e as métricas agregadas escondem tanto a transferência de falhas entre etapas (TIAN *et al.*, 2026) quanto a origem dos erros (MADEN *et al.*, 2026; MULIAN *et al.*, 2026). O descompasso entre resultados e ação descrito por Maden *et al.* (2026) tem correspondência na distribuição do mapa: pesquisadores conduzem a avaliação em 67,4% dos estudos e engenheiros em 12,3%.

5.2. Quem pode dizer que a resposta está certa

A segunda parte da hipótese da Seção 1, se o conhecimento de domínio é o gargalo da avaliação, recebe resposta parcial, porque em 21 dos 46 estudos lidos o especialista não participa e a hipótese não pode ser testada neles. A dependência do especialista aparece como

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

desafio em 15 dos 46 estudos (Quadro 1), e essa dependência inclui a escassez e, acima dela, a dificuldade de transferir o seu julgamento para um instrumento (Seção 4.7). As respostas encontradas na camada profunda são cinco: explicitar os critérios tácitos e formalizá-los em métricas computáveis validadas pelo próprio especialista (DEMYANENKO, 2026); tratá-lo como "*benchmark* vivo", em colaboração contínua (MADEN *et al.*, 2026, p. 7, tradução própria); construir camadas de rótulos com autoridade crescente (QIAN *et al.*, 2025); derivar rótulos das ações observáveis do usuário (ALBUQUERQUE *et al.*, 2024); e levar a avaliação ao usuário final quando só ele pode julgar (SAYEED; PIAL; IQBAL, 2025). Para quem desenvolve em domínio de risco, a primeira pergunta do projeto de avaliação passa a ser quem pode julgar esta resposta; a escolha da métrica vem depois.

5.3. O LLM como juiz e a validação do validador

O crescimento do LLM como juiz, de 0 em 12 estudos (2022) a 42 em 314 (2026), acompanha em parte o crescimento do próprio *corpus*. O juiz é um recurso para ganhar escala, cujo erro só é conhecido quando medido contra o julgamento humano no caso de uso específico. A própria condução desta revisão ilustra o padrão: o instrumento de triagem, validado contra 393 decisões humanas e testado com cinco modelos de três fornecedores, não atingiu a concordância de referência para triagem por uma só pessoa (tampouco a dupla humana a atingiu), e a calibração com exemplos esteve associada a maior concordância no mesmo modelo (Seção 4.8). A resposta adotada (regra validada com erro medido, auditoria humana por amostra, vereditos arquivados e exigência de nova validação a cada mudança de modelo) é a que a camada profunda recomenda para produtos (YU, 2026; GAJJAR, 2025; TAN *et al.*, 2024) e a que a posição conjunta das organizações de síntese de evidências recomenda para revisões (FLEMYNG *et al.*, 2025).

5.4. Implicações para a prática e para a pesquisa

Para a prática, os achados sugerem quatro hipóteses de trabalho, cada uma apoiada em poucos estudos da camada profunda. Primeira, a avaliação é um ciclo contínuo de calibração: critérios, conjuntos de referência e juízes automáticos perdem validade quando o *prompt* ou o modelo muda (GAJJAR, 2025; PARNIN *et al.*, 2025; YU, 2026). Segunda, nos domínios de risco o tempo do especialista é a restrição que dimensiona o resto. Terceira, as portas de verificação devem ser projetadas para falhar de forma observável, com instrumentação que atribua a falha a um componente, porque nenhum estudo lido zerou o escape (GAJJAR, 2025; GOH *et al.*, 2025; NAHAR *et al.*, 2025). Quarta, a divergência entre métricas *offline* e *online* merece medição, porque a camada profunda a reporta em poucos estudos (NAHAR *et al.*, 2025; SAH *et al.*, 2026).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Para a pesquisa, ficam três lacunas: a avaliação em produção, subdocumentada no mapa e descrita como de sinal fraco na camada profunda; o levantamento e a validação de critérios tácitos de especialistas; e o reporte padronizado da taxa de erro de avaliadores automáticos, juízes e instrumentos de triagem incluídos, para o qual os vereditos arquivados deste trabalho oferecem um conjunto de dados de partida.

5.5. Limitações e ameaças à validade

Validade de construto. A fronteira entre avaliação como contribuição e avaliação incidental, central aos critérios de elegibilidade, envolve julgamento. Ela foi operacionalizada em regras explícitas com exemplos nomeados; como a auditoria cega mostrou que os casos-limite do *corpus* são mais difíceis que os da calibração da primeira geração, o instrumento de segunda geração incorporou 30 decisões do autor como exemplos, e o erro da regra foi medido sobre 393 decisões humanas (Seção 4.8). Os 234 registros que ficaram sem resumo após a recuperação (387 menos os 153 recuperados) foram triados apenas pelo título, condição que degrada a triagem tanto do instrumento quanto do revisor humano.

Validade interna. A revisão tem revisor único. Os controles (Seção 3.3) foram fixados no protocolo ou em emendas datadas antes de sua execução. A concordância entre o autor e a segunda pessoa revisora ficou abaixo do limiar de 0,6 (Seção 4.1); a calibração da revisora usou dez registros, menos que o piloto de 50 a 100 indicado pela Cochrane (GARRITTY *et al.*, 2024), e nas discordâncias o autor foi o mais inclusivo, coerente com o privilégio à revocação da regra adotada. A diferença de mecanismo de invocação entre os modelos (Seção 3.5) impede comparação direta entre eles e deve ser lida como limitação da Tabela S1. Como a composição do *corpus* depende dos vereditos arquivados dos três modelos da regra, sua reprodução com outros modelos exige nova validação.

Validade externa. A fragmentação terminológica do campo ameaça a cobertura da busca; a resposta foi o desenho híbrido em três braços. A validação da busca contra os dez estudos de referência é parcialmente circular, porque o terceiro braço parte deles: dos dez, o primeiro braço recuperou três e o segundo recuperou cinco (com sobreposição de um estudo), e os três restantes entraram pelo terceiro braço ou diretamente como parte do conjunto de referência. A literatura cinzenta industrial fica fora do escopo, o que pode inflar a fração de estudos em cenário acadêmico e com pesquisadores como avaliadores, e reduzir a de estudos em produção.

Validade de conclusão. A classificação do mapa foi feita pelo instrumento de classificação a partir do resumo, com 164 campos reclassificados pelo autor (Seção 3.4), e validada às cegas em 40 estudos para cinco campos; os campos tipo de sistema, domínio de

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

aplicação, atributos de qualidade e métricas nomeadas não têm validação por amostra. A triagem e a validação do mapa usam convenções opostas para a incerteza: na triagem, incerteza conta como não inclusão; na validação do mapa, dúvida conta como elegível. Das 8 dúvidas na amostra de 40, 7 estão entre as 27 inclusões da regra: com dúvida contada como elegível, a falsa inclusão é de 4 em 27 (14,8%); com dúvida contada como não elegível, sobe para 11 em 27 (40,7%). A estimativa reportada na Seção 4.1 adota a dúvida como elegível; o pior caso é informado aqui. A síntese temática da camada profunda teve codificador único (o autor), sem segundo codificador.

Execução do protocolo. Nove desvios em relação ao protocolo registrado são declarados. Primeiro e segundo, os critérios I3 (idioma) e I5 (tipo de veículo) foram ampliados sem registro como emenda: I3 passou de “inglês” para “inglês ou português”, para incluir um estudo em português relevante à pergunta de pesquisa (ALBUQUERQUE *et al.*, 2024), e I5, de “arXiv” para “repositório equivalente”. Terceiro, a validação de cobertura em Scopus e IEEE Xplore, prevista no protocolo como condicional ao acesso institucional, não foi realizada. Quarto, a rodada de *snowballing* usou como sementes os 40 estudos do ensaio inicial da camada profunda (39 dos quais integram a seleção final de 46); a reexecução a partir do conjunto final, prevista na quinta emenda, não foi realizada por restrição de prazo. Quinto, os dez estudos secundários que sintetizam práticas de avaliação foram retidos no *corpus*, em desvio da exclusão prevista na terceira emenda. Sexto, as inclusões da rodada de *snowballing* não receberam as auditorias cegas prometidas na quinta emenda. Sétimo, a quinta emenda previa que um campo com *kappa* abaixo de 0,6 seria reclassificado pelo autor ou retirado da síntese; a fase do ciclo de vida (*kappa* de 0,41 na classificação fina) foi mantida porque a classificação binária que sustenta a conclusão principal teve *kappa* de 0,66, com a ressalva da Seção 4.4. Oitavo, a amostra de elegibilidade em texto integral de 15 estudos do mapa, prevista na quinta emenda, não foi executada como atividade separada; a elegibilidade foi medida nos 40 estudos da validação do mapa (Seção 4.1). Nono, a retriagem dos 182 registros sem resumo que tinham ao menos um voto de incerteza, executada pelo autor conforme a regra leniente do protocolo, identificou 130 potencialmente elegíveis pelo título; como nenhum modelo os incluiu, eles permanecem fora do *corpus* e são reportados como medida da cobertura que a regra sacrifica. O mapa contém 87 registros sem resumo cuja inclusão foi decidida pelo autor ou pela regra com ao menos um voto de inclusão; nesses, os campos que o título não permite classificar ficam como indeterminados.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

6. CONSIDERAÇÕES FINAIS

Este trabalho perguntou como a avaliação de produtos de *software* baseados em LLM é praticada ao longo do ciclo de vida, segundo a literatura publicada a partir de 2022. A resposta é um mapa de 881 estudos. A prática documentada apoia-se em estudos com usuários, experimentos controlados e avaliação humana; LLM como juiz e suítes de *benchmark* estão em ascensão, e as métricas não têm padronização. A avaliação concentra-se nas fases de desenvolvimento e pré-implantação, e a produção é a menos documentada. Entre os atributos, corretude e usabilidade dominam e os operacionais são raros. Pesquisadores conduzem a maioria das avaliações; usuários finais participam em 35,3% dos estudos e especialistas de domínio em 21,0%, com participação concentrada na saúde (atributos e domínio de aplicação são estimativas do instrumento de classificação, sem validação por amostra).

Sobre os desafios, a camada profunda mostra que a dificuldade começa antes da medição, na definição do que é uma resposta correta, e se repete em cada instrumento que tenta substituir essa definição: conjuntos de referência perecíveis, rótulos que dependem de especialistas escassos, juízes automáticos que precisam ser validados contra esses mesmos especialistas e portas de verificação que reduzem o escape sem zerá-lo. A hipótese de trabalho da Seção 1 recebe resposta parcial: o especialista de domínio participa da avaliação em um quinto dos estudos; na camada profunda, o especialista está presente em 25 dos 46 estudos e a dependência dele aparece como desafio em 15 dos 46. Para a engenharia de *software*, os estudos lidos indicam que a especificação completa do comportamento correto, pressuposto do teste tradicional, não está disponível antes da avaliação nesses produtos e resulta dela; avaliar aproxima-se de uma calibração contínua, dimensionada pelo tempo de quem pode julgar a resposta.

As contribuições são quatro. A primeira é o mapa, com as distribuições por objetivo específico, a síntese temática dos desafios e o *corpus* público depositado no projeto do protocolo. A segunda é o relato metodológico, inspecionável registro a registro, de uma triagem instrumentada por modelos de linguagem, com calibração, validação contra decisões humanas, auditoria cega e teste de cinco modelos de três fornecedores sob o mesmo *prompt*. A terceira é a evidência, obtida na própria condução, de que nem os modelos isolados nem a dupla humana atingiram a concordância de referência para triagem por revisor único, que todos divergiram do autor na mesma direção e que a calibração com exemplos esteve associada a maior concordância no mesmo modelo, em amostra que sobrerrepresenta os casos-limite (Seção 4.8). A quarta é a ferramenta de triagem e verificação descrita na Seção 3.3, disponível no projeto do protocolo. O alcance dessas conclusões é o delimitado na Seção 5.5: um mapa construído por

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

resumo, com 80,7% dos estudos incluídos por regra validada cujo erro de triagem foi medido na busca principal e cuja falsa inclusão no mapa foi estimada em amostra de 40 estudos (Seção 4.1), e leitura integral restrita a uma camada de estudos centrais.

6.1. Financiamento

Este trabalho não recebeu financiamento de agência de fomento nem de empresa. Os custos de execução foram pagos pelo primeiro autor.

6.2. Conflitos de interesse

O primeiro autor desenvolve uma plataforma de análise baseada em LLM, citada na Introdução; os autores declaram não haver outros conflitos de interesse. Os autores não têm vínculo com os fornecedores dos modelos utilizados.

6.3. Declaração de uso de inteligência artificial

Modelos de linguagem foram usados como instrumentos do método: na triagem, na classificação da camada larga e no pré-preenchimento das fichas da camada profunda, com modelos, versões, *prompts* e validação descritos nas Seções 3.3 a 3.5. Os autores respondem integralmente pelo conteúdo publicado.

REFERÊNCIAS

ALBUQUERQUE, A. M. de *et al.* Avaliação de Aplicações de Geração Aumentada de Recuperação por meio de Feedback Implícito. In: *Anais Estendidos do Simpósio Brasileiro de Banco de Dados (SBBD)*. Porto Alegre: SBC, 2024. p. 253–259. DOI: 10.5753/sbbd_estendido.2024.243903.

ARKSEY, H.; O'MALLEY, L. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology*, v. 8, n. 1, p. 19–32, 2005. DOI: 10.1080/1364557032000119616.

COSCIA, A. J.; ENDERT, A. KnowledgeVIS: Interpreting Language Models by Comparing Fill-in-the-Blank Prompts. *IEEE Transactions on Visualization and Computer Graphics*, v. 30, n. 9, p. 6520–6532, 2024. DOI: 10.1109/tvcg.2023.3346713.

DEMYANENKO, N. *From Tacit Expertise to Computable Criteria: A Knowledge Elicitation Framework for LLM-Based Evaluation of Complex Multi-Modal Content*. 2026. Zenodo. DOI: 10.5281/zenodo.20395370.

FLEMYNG, E. *et al.* Position statement on artificial intelligence (AI) use in evidence synthesis across Cochrane, the Campbell Collaboration, JBI and the Collaboration for Environmental Evidence 2025. *Cochrane Database of Systematic Reviews*, 2025. Art. ED000178. DOI: 10.1002/14651858.ED000178.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

GAJJAR, U. AI-Augmented Testing solution of LLM-Integrated Mobile Applications: Architecture, Test Oracle Strategies, and Empirical Evaluation. *International Journal of Technology & Emerging Research*, v. 1, n. 6, p. 67–73, 2025. DOI: 10.64823/ijter.2506008.

GARCIA, A. A. *et al.* Red Teaming LLMs as Socio-Technical Practice: From Exploration and Data Creation to Evaluation. In: *CHI Conference on Human Factors in Computing Systems*. New York: ACM, 2026. p. 1–18. DOI: 10.1145/3772318.3790792.

GARRITTY, C. *et al.* Updated recommendations for the Cochrane rapid review methods guidance for rapid reviews of effectiveness. *BMJ*, v. 384, p. e076335, 2024. DOI: 10.1136/bmj-2023-076335.

GEBREEGZIABHER, S. A. *et al.* MetricMate: An Interactive Tool for Generating Evaluation Criteria for LLM-as-a-Judge Workflow. In: *Annual Symposium on Human-Computer Interaction for Work*. New York: ACM, 2025. p. 1–18. DOI: 10.1145/3729176.3729199.

GOH, J. Y. *et al. Measuring What Matters: A Framework for Evaluating Safety Risks in Real-World LLM Applications*. 2025. arXiv. DOI: 10.48550/arXiv.2507.09820.

GUO, G. *et al.* MORTAR: Multi-Turn Metamorphic Testing for LLM-Based Dialogue Systems. *IEEE Transactions on Software Engineering*, v. 52, n. 9, p. 2620–2637, 2026. DOI: 10.1109/tse.2026.3701230.

HE, J. *et al.* POEM: Interactive Prompt Optimization for Enhancing Multimodal Reasoning of Large Language Models. In: *IEEE Pacific Visualization Conference*. [S.l.]: IEEE, 2025. p. 36–46. DOI: 10.1109/pacificvis64226.2025.00010.

JIANG, E. *et al.* PromptMaker: Prompt-based Prototyping with Large Language Models. In: *CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA)*. New York: ACM, 2022. p. 1–8. DOI: 10.1145/3491101.3503564.

KHRAISHA, Q. *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*, v. 15, n. 4, p. 616–626, 2024. DOI: 10.1002/jrsm.1715.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing Systematic Literature Reviews in Software Engineering*. Keele: Keele University; Durham: Durham University, 2007. (EBSE Technical Report, EBSE-2007-01).

LAURO, Q. R.; BIGHAM, J. P.; KOTTURI, Y. Exploring the Role of Social Support When Integrating Generative AI in Small Business Workflows. In: *ACM Conference on Computer Supported Cooperative Work*. New York: ACM, 2024. p. 485–492. DOI: 10.1145/3678884.3681895.

LEVAC, D.; COLQUHOUN, H.; O'BRIEN, K. K. Scoping studies: advancing the methodology. *Implementation Science*, v. 5, n. 1, p. 69, 2010. DOI: 10.1186/1748-5908-5-69.

LI, H. *et al. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods*. 2024. arXiv. DOI: 10.48550/arXiv.2412.05579.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

LIU, L. *et al.* Counterfactual Evaluation for Blind Attack Detection in LLM-based Evaluation Systems. In: *International Joint Conference on Natural Language Processing and Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*. Stroudsburg: ACL, 2025. p. 572–584. DOI: 10.18653/v1/2025.ijcnlp-long.33.

LIU, M. X. *et al.* "We Need Structured Output": Towards User-centered Constraints on Large Language Model Output. In: *CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA)*. New York: ACM, 2024. p. 1–9. DOI: 10.1145/3613905.3650756.

LÓPEZ-PINEDA, A. *et al.* Validation of large language models (Llama 3 and ChatGPT-4o mini) for title and abstract screening in biomedical systematic reviews. *Research Synthesis Methods*, v. 16, n. 4, p. 620–630, 2025. DOI: 10.1017/rsm.2025.15.

MADEN, W. van der *et al.* Results-Actionability Gap: Understanding How Practitioners Evaluate LLM Products in the Wild. In: *CHI Conference on Human Factors in Computing Systems*. New York: ACM, 2026. p. 1–17. DOI: 10.1145/3772318.3791069.

MAGALHÃES, C. V. C. de *et al.* Testing the Untestable? An Empirical Study on the Testing Process of LLM-Powered *Software* Systems. In: *IEEE International Working Conference on Source Code Analysis and Manipulation*. [S.l.]: IEEE, 2025. p. 78–88. DOI: 10.1109/scam67354.2025.00015.

MAILACH, A. *et al.* Themes of Building LLM-Based Applications for Production: A Practitioner's View. In: *IEEE/ACM International Conference on AI Engineering – Software Engineering for AI*. [S.l.]: IEEE, 2025. p. 18–30. DOI: 10.1109/cain66642.2025.00011.

MULIAN, H. *et al.* AgentFixer: From Failure Detection to Fix Recommendations in Agentic Systems. In: *International Workshop on Agentic Engineering*. New York: ACM, 2026. p. 96–103. DOI: 10.1145/3786167.3788427.

NAHAR, N. *et al.* Beyond the Comfort Zone: Emerging Solutions to Overcome Challenges in Integrating LLMs into *Software* Products. In: *IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. [S.l.]: IEEE, 2025. p. 516–527. DOI: 10.1109/icse-seip66354.2025.00051.

NAVARRO, K. F.; SYRIANI, E.; ARAWJO, I. Reporting and Reviewing LLM-Integrated Systems in HCI: Challenges and Considerations. In: *CHI Conference on Human Factors in Computing Systems*. New York: ACM, 2026. p. 1–18. DOI: 10.1145/3772318.3790439.

PAN, R. *et al.* Tangent: An Empirical Study of Testing Practices for LLM-Based Agent Applications. In: *IEEE/ACM International Conference on Automated Software Engineering*. New York: ACM, 2026. p. 1483–1495. DOI: 10.1145/3832783.3837414.

PARNIN, C. *et al.* Building Your Own Product Copilot: Challenges, Opportunities, and Needs. In: *IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*. [S.l.]: IEEE, 2025. p. 338–348. DOI: 10.1109/saner64311.2025.00039.

PETERSEN, K. *et al.* Systematic mapping studies in *software* engineering. In: *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. Swindon: BCS Learning & Development, 2008. DOI: 10.14236/ewic/ease2008.8.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in *software* engineering: An update. *Information and Software Technology*, v. 64, p. 1–18, 2015. DOI: 10.1016/j.infsof.2015.03.007.

POTTERS, N.; HOFMAN, T. *Auto-DSM Under the Lens: A Black-Box Evaluation Framework for LLM-Based DSM Generation*. 2026. arXiv. DOI: 10.48550/arXiv.2607.05985.

PRIEM, J.; PIWOWAR, H.; ORR, R. *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts*. 2022. arXiv. DOI: 10.48550/arXiv.2205.01833.

QIAN, C. *et al.* LLM Adoption in Data Curation Workflows: Industry Practices and Insights. In: *CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA)*. New York: ACM, 2025. p. 1–10. DOI: 10.1145/3706599.3719677.

SAH, C. K. *et al.* Fairness Evaluation in LLM-Based Educational Recommender Systems: A Framework for Equitable Learning. In: *Artificial intelligence*. [S.l.]: IntechOpen, 2026. DOI: 10.5772/intechopen.1012999.

SAYEED, S.; PIAL, S. M. R.; IQBAL, A. NuevAI: Streamlined Dataset Generator and Human Evaluator System for Development of Pedagogical Conversational Agents. In: *Biannual Conference of the Italian SIGCHI Chapter*. New York: ACM, 2025. p. 1–11. DOI: 10.1145/3750069.3750094.

SHANKAR, S. *et al.* Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. In: *ACM Symposium on User Interface Software and Technology (UIST)*. New York: ACM, 2024. p. 1–14. DOI: 10.1145/3654777.3676450.

SHANKAR, S.; HUSAIN, H. *Evals for AI Engineers*: Systematically Measuring and Improving AI Applications. Santa Rosa: O'Reilly Media, 2026. ISBN 979-8-341-66072-4.

SORDO, A. *et al. STELLAR-E: a Synthetic, Tailored, End-to-end LLM Application Rigorous Evaluator*. 2026. arXiv. DOI: 10.48550/arXiv.2604.24544.

SZYMANSKI, A. *et al.* Key Considerations for Domain Expert Involvement in LLM Design and Evaluation: An Ethnographic Study. In: *ACM International Conference on Intelligent User Interfaces*. New York: ACM, 2026. p. 617–630. DOI: 10.1145/3742413.3789105.

TAN, T. F. *et al. Fine-tuning Large Language Model (LLM) Artificial Intelligence Chatbots in Ophthalmology and LLM-based evaluation using GPT-4*. 2024. arXiv. DOI: 10.48550/arXiv.2402.10083.

TIAN, H. *et al.* RAGExplorer: A Visual Analytics System for the Comparative Diagnosis of RAG Systems. *IEEE Transactions on Visualization and Computer Graphics*, v. 32, n. 6, p. 4807–4817, 2026. DOI: 10.1109/tvcg.2026.3694443.

TRICCO, A. C. *et al.* PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine*, v. 169, n. 7, p. 467–473, 2018. DOI: 10.7326/M18-0850.

VANEL, L. *et al.* Socio-Emotional Response Generation: A Human Evaluation Protocol for LLM-Based Conversational Systems. In: *International Conference on Affective Computing and*

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Intelligent Interaction Workshops and Demos. [S.l.]: IEEE, 2024. p. 290–305. DOI: 10.1109/aciiw63320.2024.00057.

VERO, M. *et al. Honeyval: A Comprehensive Evaluation Framework for LLM-powered HTTP Honeypots*. 2026. arXiv. DOI: 10.48550/arXiv.2605.29963.

WANG, S. *et al. Configurable multi-agent framework for scalable and realistic testing of llm-based agents*. 2025. arXiv. DOI: 10.48550/arXiv.2507.14705.

WOHLIN, C. Guidelines for snowballing in systematic literature studies and a replication in *software* engineering. In: *International Conference on Evaluation and Assessment in Software Engineering (EASE)*. New York: ACM, 2014. p. 1–10. DOI: 10.1145/2601248.2601268.

YEHUDAI, A. *et al.* A Survey on Evaluation of LLM-based Agents. In: *Findings of the Association for Computational Linguistics: ACL 2026*. Stroudsburg: ACL, 2026. p. 26690–26714. DOI: 10.18653/v1/2026.findings-acl.1330.

YU, L. *et al.* Measuring the quality of generative AI systems: Mapping metrics to quality characteristics. *Information and Software Technology*, v. 186, p. 107802, 2025. DOI: 10.1016/j.infsof.2025.107802.

YU, Y. *Evaluating the Evaluator: A Protocol for Measuring Evaluator Dependence in LLM-Based Safety Assessment*. 2026. figshare. DOI: 10.6084/m9.figshare.32886083.v1.

ZAMFIRESCU-PEREIRA, J. *et al.* Herding AI Cats: Lessons from Designing a Chatbot by Prompting GPT-3. In: *Designing Interactive Systems Conference*. New York: ACM, 2023. p. 2206–2220. DOI: 10.1145/3563657.3596138.

ZHAO, Z. *et al.* On the Workflows and Smells of Leaderboard Operations (LBOps): An Exploratory Study of Foundation Model Leaderboards. *IEEE Transactions on Software Engineering*, v. 51, n. 4, p. 929–946, 2025. DOI: 10.1109/tse.2025.3533972.

MATERIAL SUPLEMENTAR

O material suplementar está disponível no projeto do protocolo (https://osf.io/xygec). O Material suplementar S1 traz a validação do instrumento de triagem, com a Tabela S1. O Material suplementar S2 traz a Tabela S2 (contagem de estudos por método de avaliação e fase do ciclo de vida), a Figura S1 (estudos do *corpus* por ano de publicação e situação do veículo) e a Figura S2 (temas de desafios da camada profunda e relações entre eles).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.