

REVISÃO POR PARES · AGENTES DE IA

Revisão por pares assistida por agentes de IA



Análise técnica e crítica do Multi-Layered Review (MLR) da
Sakana AI: o que a manchete de 73% realmente mede

Aridio Silva · Pesquisador Independente · @aridiosilva

ORCID 0009-0008-2411-6995 · 13-Outubro-2026

Roteiro: a estrutura do paper

1

Introdução

2

Verificação da alegação

3

Fundamentação

4

Estado da arte

5

O benchmark

6

O sistema MLR

7

Protocolo experimental

8

Resultados

9

Evidência independente

10

Discussão e credibilidade

11

Limitações

12

Lacunas

13

Perspectivas

14

Requisitos

15

Conclusão

Resumo executivo



73,43%

contradições sintéticas
na alegação principal (4
revisões)



40,95%

em todas as distâncias do
benchmark (4 revisões)



16,11%

motivo real de
retratação,
correspondência exata



0

medições de precisão
(falsos alarmes) no paper

Por que importa

Detectar erros é a função mais cara da revisão
por pares; o MLR mira exatamente isso.

Implicações positivas

- ◆ Forte em erros plantados na alegação central
- ◆ Custo baixo: cerca de US\$ 0,47 por revisão

Implicações negativas

- Cai a 16–26% em erros reais
- Falsos alarmes não foram medidos

O problema: revisão por pares sob pressão

- Mais submissões, revisões de menor qualidade (Teo et al., 2026)
- ▲ NeurIPS: cerca de metade dos aceitos por um comitê seria rejeitada por outro
- ◆ LLMs já em uso: estudo na NEJM AI, ensaio do ICLR 2025 (20 mil revisões) e piloto do AAAI-26 (22.977 submissões)
- ▶ Lacuna: avaliar por imitação de revisões humanas, não por achar erros

Por que importa

Detectar erros é a tarefa mais custosa da revisão, e a mais carente de braços.

Implicações positivas

- ◆ Escalar a triagem de erros
- ◆ Liberar tempo do revisor humano

Implicações negativas

- Risco de delegar sem ler
- Notas inflacionadas em uso indiscriminado

Quatro questões de pesquisa

Q1 O que a manchete afirma e o que o paper realmente mede?

Q2 Como foram construídos o benchmark, o sistema e o protocolo?

Q3 Quais resultados se sustentam e quais dependem de condições favoráveis?

Q4 Que lacunas e ameaças à validade permanecem?

Por que importa

Perguntas claras separam fato de manchete.

Implicações positivas

- ◆ Verificação rastreável até a fonte
- ◆ Foco em evidência, não em opinião

Implicações negativas

- Escopo limitado ao paper e a fontes abertas
- Parte do apêndice não foi lida

A manchete frente ao que o paper mede

Aspecto	Manchete	Paper (Teo et al., 2026)
Quem detecta	"3 agentes"	Só o Review Agent (Claude Sonnet 4)
Quantas revisões	Implícito: uma	Quatro; basta uma acertar
Tipo de erro	"Erros"	Contradições sintéticas (GPT-4.1)
Gravidade	"Alegações centrais"	Distância 0 no grafo de conhecimento
Base de dados	"Artigos"	257 papers de aprendizado de máquina
Total do benchmark	Não citado	40,95% (4 revisões) e 28,32% (1 revisão)
Erros reais	Não citados	26,07% (similar) e 16,11% (exato)

Por que importa

A manchete mistura configurações e esconde o recorte medido.

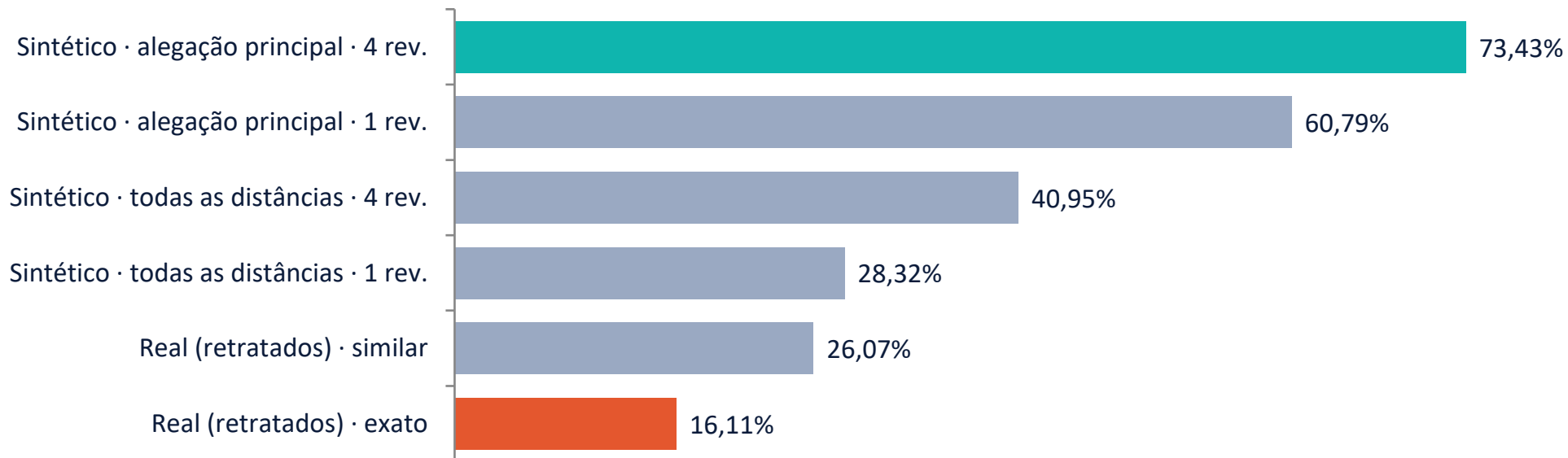
Implicações positivas

- ◆ 73,43% é real e reportado no paper
- ◆ Resultado verificável

Implicações negativas

- Leitura apressada superestima o sistema
- A imprensa repete a manchete

O 73% cai para 16% quando os erros são reais



Acerto do MLR por tipo de erro e configuração (%). Fonte: Teo et al. (2026), Tabelas 3 e 14.

Por que importa

O mesmo sistema, seis medições: o número depende do tipo de erro e do ensemble.

Implicações positivas

- ◆ Sintético: 73,43% na alegação principal
- ◆ Uma só revisão já acerta 60,79%

Implicações negativas

- Erro real exato: apenas 16,11%
- Ensemble multiplica o custo por 4

Método e legenda de verificação

- ◆ Confirmado no paper: número lido no texto do paper ou em trecho dele
- ▲ Fonte secundária: cobertura jornalística ou blog, a conferir
- ✦ Inferência própria: cálculo ou interpretação do autor desta análise
- Lido: texto principal, apêndices A.1 a A.4 e trechos de A.5, A.6 e B.1
- Não lido: OpenReview (TMLR), Substack original, Figura 28, A.5, A.7, B.2 a B.4

Por que importa

Rastrear cada número até a fonte evita repetir manchetes.

Implicações positivas

- ◆ Rigor de fontes e datas
- ◆ Limites declarados

Implicações negativas

- OpenReview bloqueado: revisões do TMLR não lidas
- Substack original não localizado

Conceitos centrais

Assistência × substituição

A IA apoia o revisor humano; não o substitui

Grafo e distância

Distância até a alegação principal mede a severidade

LLM como juiz

Um modelo decide se o erro foi achado

Precisão e recall

Recall: erros achados; precisão: apontamentos corretos

Contradição sintética

Erro lógico inserido de forma controlada

Entender antes de criticar

Leitura estruturada antes da revisão (Keshav)

Detecção em ≥ 1 de k

Conta se qualquer revisão acertar (como pass@k)

Injeção de prompt

Instrução oculta que tenta manipular o revisor

Por que importa

Vocabulário comum evita confundir recall com precisão.

Implicações positivas

- ◆ Leitura crítica de números
- ◆ Base para comparar sistemas

Implicações negativas

- Termos novos elevam a barreira
- Definições variam entre artigos

Quatro fases, de 2007 a 2026

1. Primeiros testes

2007–2023

Keshav (2007)
Yuan et al. (2022)
Liu & Shah (2023)
Robertson (2023)

2. Escala e agentes

2024

Liang et al. (NEJM AI)
MARG
AgentReview
AI Scientist
Reviewer2
WithdrarXiv
Skarlinski et al.
Ye et al.

3. Verificação

2025

DeepReview
ReviewAgents
GAR
SPOT
Zhang & Abernethy
FLAWS
Thakkar et al.
Hidden Prompts

4. Implantação

2026

Piloto AAI-26
ReviewGrounder
PRISM
ReviewBench
PeerReview Bench
MLR (Sakana AI)
AI Scientist (Nature)

Por que importa

Entender a trajetória mostra o que o MLR de fato acrescenta.

Implicações positivas

- ◆ Campo maduro: de testes a pilotos em escala
- ◆ Reuso de ideias (Three-Pass, grafos)

Implicações negativas

- Ritmo rápido dificulta replicar
- Parte da literatura sem autoria verificada

Do teste exploratório à implantação

- 2022–2023: modelos geram feedback e detectam erros inseridos
- ▶ 2024: utilidade medida em larga escala, sistemas multiagente e o AI Scientist
- ▲ 2025: raciocínio profundo, benchmarks de erro e textos ocultos em manuscritos
- ◆ 2026: implantação (AAAI-26), avaliação por especialistas e o MLR
- O revisor do AI Scientist original extrai texto do PDF sem ler figuras e tabelas

Por que importa

O MLR responde a limites conhecidos: envia o PDF como texto e imagem.

Implicações positivas

- ◆ Acúmulo de evidência desde 2022
- ◆ Piloto real em 22.977 submissões

Implicações negativas

- Utilidade percebida ≠ taxa de falhas achadas
- Notas inflacionadas e uso não declarado

Sistemas de revisão por IA

Sistema	Abordagem	Evidência disponível
MLR (Teo et al., 2026)	Cadeia de 3 passos + agentes auxiliares (Claude)	Seção 8 deste trabalho
AI Scientist Reviewer	Ensemble + meta-revisão + reflexão	Lê texto do PDF, sem figuras
AgentReview	Simula revisores, autores e AC	Baseline do paper
MARG	Agente líder + especialistas	Sem comparação direta
GAR	Grafo do paper + 3 a 6 revisores	Previsão de aceitação
Reviewer2	Prompts por aspecto + geração	Descrito no paper
ReviewAgents	Cadeias de raciocínio; agentes treinados	Ablação do nº de agentes
DeepReview	Raciocínio estruturado	DeepReview-13K / Bench
ReviewGrounder	Rubricas + ferramentas	ReviewBench
Reviewer3.com	Multiagente comercial	Comparado no ReviewBench
Stanford · OpenAIReview · CMU	Plataformas de revisão agêntica	P/R/F1 no PeerReview Bench

Por que importa

Campo cresceu em geração, avaliação e verificação de erros.

Implicações positivas

- ◆ Muitas abordagens para comparar
- ◆ Plataformas abertas já disponíveis

Implicações negativas

- Poucas comparações diretas entre elas
- Modelos mudam, resultados envelhecem

Benchmarks de detecção de erros

Benchmark	Tipo de erro	Escala	Avaliação
Contradiction Benchmark (Teo)	Sintético, grafo com severidade	1.164 em 257 papers de ML	LLM-juiz (o3)
WithdrarXiv-Check	Real, retratações	211 papers	LLM-juiz (o3)
FLAWS (Xi et al.)	Sintético, invalida alegações	713 pares paper-erro	Automática
SPOT (Son et al.)	Real, verificação automática	Não verificada	Não verificada
PeerReview Bench (Kim et al.)	Real, itens validados por especialistas	78 papers	Meta-revisor de IA
ContraDetect (Skarlinski)	Contradições entre artigos	Não verificada	Especialistas

Por que importa

O benchmark certo define o que 'bom' significa para um revisor automático.

Implicações positivas

- ◆ Erro sintético dá alvo inequívoco
- ◆ Escala sem anotação manual

Implicações negativas

- Sintético pode ser mais fácil que real
- Tarefas não são comparáveis entre si

Posicionamento do MLR

- Único benchmark da lista que gradua a severidade do erro por distância no grafo
- ▶ Outros trabalhos de 2026 avaliam por rubricas, dimensões ou processo
- ▲ Estudo com agentes Claude em análise de dados: bom em erros de implementação, fraco em crítica de domínio
- ✦ IA e humanos acham problemas largamente distintos: tese de complementaridade

Por que importa

Posicionar o MLR evita comparar tarefas diferentes como se fossem iguais.

Implicações positivas

- ◆ Complementa o revisor humano
- ◆ Foco em verificação, não em imitação

Implicações negativas

- Cobre ML; outros domínios não testados
- Perfil 'irregular' entre tipos de erro

Como o benchmark é construído

1. Coleta

257 papers de ACL, AISTATS, CVPR, ICML (2025) e NeurIPS (2024) com LaTeX do arXiv

licenças CC

2. Grafo

Alegações, evidências, métodos e implementações com citação literal

Gemini 2.5 Pro

3. Contradição

Um nó por distância; o parágrafo é reescrito para contradizer

GPT-4.1

4. Recompilação

LaTeX modificado gera o PDF: 1.164 contradições

LaTeX → PDF

5. Avaliação

O juiz compara a revisão com o parágrafo original e o modificado

o3 · 10 repetições

Por que importa

Cada contradição nasce de um nó do grafo e tem severidade conhecida.

Implicações positivas

- ◆ Alvo inequívoco e escalável
- ◆ Severidade graduada por distância

Implicações negativas

- Nem todo LaTeX recompila
- Erro de LLM pode soar artificial

Grafo de conhecimento e distância

Tipo de nó	Subtipos / definição
Claim	Principal, secundária, terciária
Evidence	Alta, média, baixa
Methodology	Técnica ou abordagem nova
Implementation	Detalhes experimentais
Mathematical Theory	Teoremas novos e demonstrados

Distância 0 = alegação principal. Resultado: 1.164 contradições (distâncias 0 a 8); análise em 0 a 6. A relação 'Contradicts' aparece em apenas 0,4% do grafo.

Por que importa

A distância transforma 'quão grave é o erro' em um número.

Implicações positivas

- ◆ Compara erros de mesmo peso
- ◆ Revela queda por severidade

Implicações negativas

- Severidade presumida, não medida
- Poucas amostras além de d6

O juiz (o3) e sua validação

Juiz	Outlines (%)	Revisões (%)
o3 (escolhido)	97,42	99,90
o4-mini	94,70	99,52
GPT-4.1	92,63	100,00
GPT-5.2	88,77	96,35
Gemini 2.5 Pro	72,40	93,11
Gemini 3 Pro	99,87	99,92



86,8%

sensibilidade do juiz em 50 acertos verificados



≈13%

dos acertos reais podem ser perdidos: taxas conservadoras

Acurácia em papers sem erro inserido (26 papers, 130 contradições). Fonte: Teo et al. (2026), Tabela 11.

Por que importa

Um juiz mal calibrado distorce todos os resultados do benchmark.

Implicações positivas

- ◆ Juiz validado em papers sem erro
- ◆ 10 repetições reduzem ruído

Implicações negativas

- Juiz de outra família que os sistemas
- Viés cruzado não discutido

Limitações do benchmark

- ▲ Contradições geradas por LLM podem soar artificiais
- Auditoria de 50 casos: 34% sem coerência com as frases vizinhas e 8% obviamente geradas por IA
- Os autores veem isso como facilitador, não como dificultador
- ✦ Um sistema adversarial poderia comparar o PDF editado com o original no arXiv

Por que importa

Mostra até onde o 73% pode ser generalizado para erros naturais.

Implicações positivas

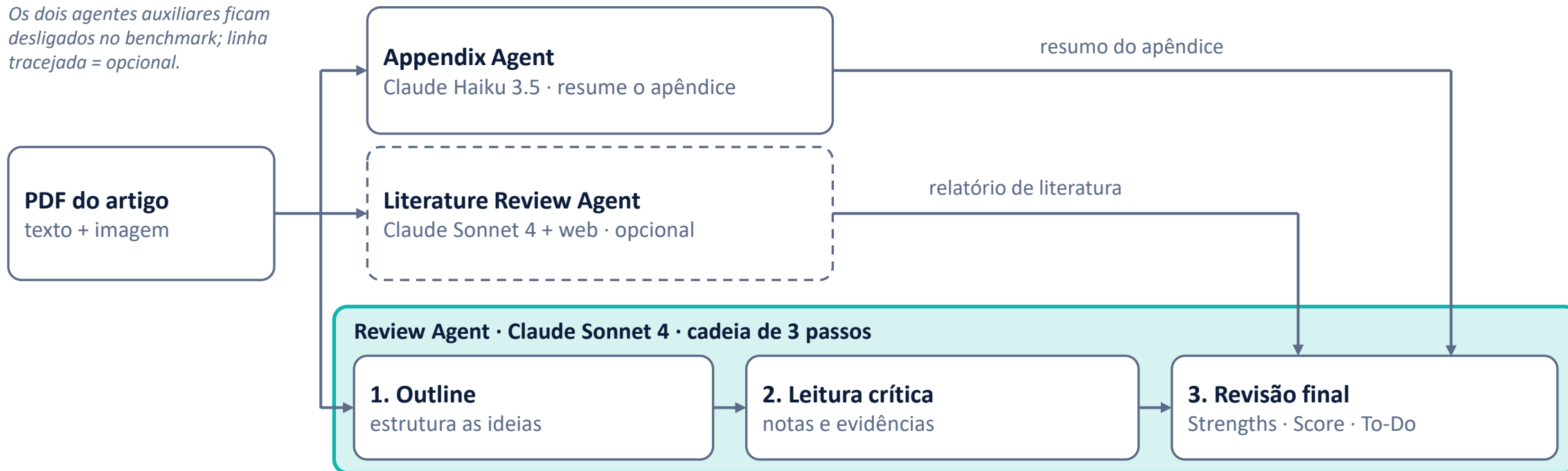
- ◆ Limitações assumidas pelos autores
- ◆ Auditoria manual do próprio conjunto

Implicações negativas

- Recall em uso real tende a ser menor
- Possível vazamento por busca no arXiv

Arquitetura e fluxo do MLR

Os dois agentes auxiliares ficam desligados no benchmark; linha tracejada = opcional.



Por que importa

Só o Review Agent foi medido no benchmark; os outros dois são auxiliares.

Implicações positivas

- ◆ PDF em texto e imagem preserva fórmulas
- ◆ Apêndice resumido evita críticas já respondidas

Implicações negativas

- Texto principal limitado a 10 páginas
- Resumo do apêndice por modelo menor perde detalhes

Papéis, modelos e artefatos

Componente	Modelo	Função
Appendix Agent	Claude Haiku 3.5	Resume o apêndice; gestão de contexto
Literature Review Agent (opcional)	Claude Sonnet 4 + busca na web	Avalia a novidade e supre o limite de conhecimento
Review Agent	Claude Sonnet 4	Cadeia de 3 prompts sobre o texto principal

Artefatos gerados

Outline do artigo

Notas críticas

Revisão final + To-Do

Resumo do apêndice

Relatório de literatura

Por que importa

Modelos comerciais prontos, sem ajuste fino: fácil de reproduzir e de trocar.

Implicações positivas

- ◆ Sem GPU nem treinamento
- ◆ Cinco artefatos úteis ao autor

Implicações negativas

- Dependência de APIs de terceiros
- Versões de modelo envelhecem

A cadeia de três passos

1

1. Outline

Estrutura as ideias centrais do artigo em alto nível



2

2. Leitura crítica

Notas com evidências, suposições, fraquezas e lacunas; corrige o 1º passo



3

3. Revisão final

Strengths, Weaknesses, Questions, Recommendation, Score (1–10) e To-Do

Por que importa

Inspirada no Three-Pass de Keshav: compreender antes de criticar.

Implicações positivas

- ◆ Ganho forte em erros distantes da alegação
- ◆ Notas intermediárias auditáveis

Implicações negativas

- Só a rubrica de nota é detalhada no prompt
- Maior custo que um prompt único

Como os multiagentes operam

- Coordenação sequencial e sem debate entre agentes
- ▶ Dois agentes auxiliares produzem insumos para um único agente revisor
- ▲ Sem papéis de autor ou area chair, ajuste fino ou raciocínio extra
- ◆ Inferência própria: 'multiagente' é generoso; a parte analítica é uma cadeia de prompts
- ◆ Variante de prompt único: custo cerca de 1/3, queda de cerca de 3,5 pontos no total

Por que importa

Mostra o que realmente gera o resultado: o modelo e a cadeia, não a 'orquestra'.

Implicações positivas

- ◆ Simples de implementar e depurar
- ◆ Variante barata disponível

Implicações negativas

- Rótulo 'multiagente' pode confundir
- Sem debate para reduzir falsos alarmes

Custo por revisão do MLR

Componente	Entrada (tokens)	Saída (tokens)	Custo (US\$)
MLR (apêndice + revisão)	189.062	3.913	0,42 + 0,05
Appendix Agent	66.521	1.011	0,05 + <0,01
Review Agent	122.541	2.902	0,37 + 0,04
Literature Agent (opcional)	312.767	2.032	0,94 + 0,03



≈ US\$ 0,47

por revisão, sem o agente de literatura



≈ US\$ 1,9

para 4 revisões (cálculo próprio)

Fonte: Teo et al. (2026), Tabela 9. Preços em US\$ por milhão de tokens: Sonnet 3/15; Haiku 0,8/4.

Por que importa

Custo baixo viabiliza segunda leitura em todo manuscrito.

Implicações positivas

- ◆ Cerca de US\$ 0,47 por artigo
- ◆ Metade dos tokens do AI Reviewer

Implicações negativas

- O ensemble de 4 multiplica por ~4
- Literatura soma ~US\$ 0,97

Mapa das avaliações: o que foi testado onde

Avaliação	Amostra	Configuração do MLR	Métrica
Contradiction Benchmark	1.164 contradições, 257 papers	Só Review Agent	Acerto; o3 × 10
WithdrarXiv-Check	211 de 245 papers (≤ 30 págs.)	Só Review Agent	Motivo da retratação; o3
Foco da revisão	Expert Review Dataset	Appendix + Review (web)	KL com humanos
Submissões de conferência	ICML 2025 · ICLR 2025 · NeurIPS 2024	Appendix + Review (web)	Correlação de notas
Manipulação	50 papers rejeitados	Só Review Agent	Variação de nota
Estudo com usuários	38 sessões	Três agentes	Likert; aceitação

Por que importa

O sistema de três agentes só foi avaliado em conjunto no estudo com usuários.

Implicações positivas

- ◆ Várias avaliações complementares
- ◆ Configurações declaradas

Implicações negativas

- 73% vem do agente isolado, não dos três
- Sistema completo sem taxa de detecção

Baselines e uma pergunta em aberto

- LLM-Review: GPT-4.1, 10 primeiras páginas, texto truncado em 6.500 tokens
- ▶ AI Reviewer: o4-mini, ensemble de 5, meta-revisão e até 5 rodadas de reflexão
- ▲ AgentReview: GPT-4o, 3 revisões, só a primeira fase
- ✦ MLR medido com 4 revisões, entre os ensembles de 3 e 5 das baselines
- Em aberto: como o paper trata páginas além da 10ª no WithdrarXiv (papers até 30 págs.)?

Por que importa

Baselines diferentes em modelo e design tornam a comparação parcialmente confundida.

Implicações positivas

- ◆ Baselines conhecidas e acessíveis
- ◆ Ablação com Claude no LLM-Review

Implicações negativas

- Erros profundos podem ficar fora das 10 págs.
- Ensembles de tamanhos diferentes

Detecção por distância (Tabela 14 do paper)

Acerto (%)	d0	d1	d2	d3	d4	d5	d6	Total
MLR, 4 revisões	73,43	47,17	31,67	27,41	21,46	15,21	38,89	40,95
MLR, 1 revisão	60,79	32,43	19,56	14,47	12,85	4,17	16,67	28,32
MLR-Combined (subconjunto)	56,67	31,42	18,03	9,89	3,77	1,43	0,00	24,81
LLM-Review + Claude	35,40	22,75	7,77	7,68	6,99	8,33	0,00	16,43
LLM-Review (GPT-4.1)	14,56	7,89	3,63	3,03	2,20	2,29	0,00	6,39
AgentReview	14,81	6,33	3,82	2,98	1,30	0,00	0,00	5,95
AI Reviewer	11,17	7,73	6,89	3,73	1,71	2,08	0,00	6,50

d7 e d8: 0,00 em todos, exceto AI Reviewer (14,00 em d7). d6 tem só 18 amostras.

Por que importa

O 73% é a fatia mais favorável: o total do benchmark é 40,95%.

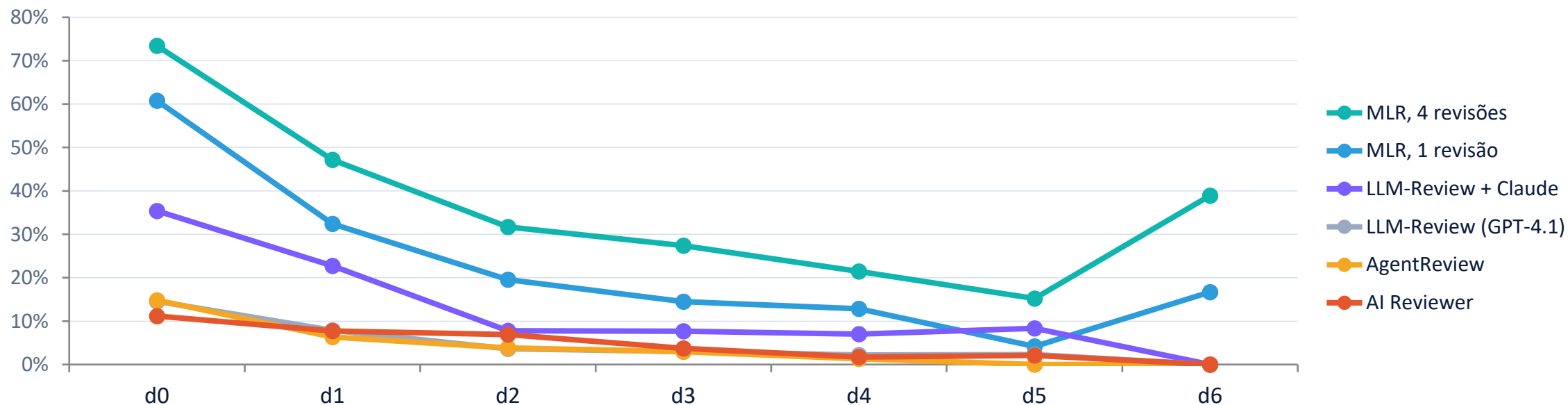
Implicações positivas

- ◆ ≈ 6,3× o melhor baseline no total
- ◆ Queda monótona calibra a severidade

Implicações negativas

- Erros a 4 passos: só 21,46%
- Poucas amostras em d5 e d6

A detecção cai de 73% para 21% do d0 ao d4



Acerto (%) por distância até a alegação principal. d6 tem 18 amostras. Fonte: Teo et al. (2026), Tabela 14.

Por que importa

O ensemble eleva a curva, mas não muda sua forma: erros distantes escapam mais.

Implicações positivas

- ◆ MLR supera as baselines GPT em todas as distâncias
- ◆ 4 revisões somam ~12 pontos

Implicações negativas

- Erros periféricos têm baixo recall
- d6 é ruído amostral (18 itens)

Quatro leituras dos números

- O 73% é a fatia mais favorável: no total, 40,95%; cerca de 6,3× a melhor baseline
- ▶ Incerteza: com ~250 itens em d0, o intervalo de 95% é de ± 5 pontos
- ▲ Ensemble: $P4 = 1 - (1 - 0,6079)^4 \approx 0,976$ se fosse independente; observado 0,7343: itens sistematicamente difíceis
- ✦ Juiz: recall real $\lesssim 73,43 \div 0,868 \approx 84,6\%$, limite superior grosseiro

Por que importa

Cálculos próprios mostram o que o ensemble e o juiz fazem com a medição.

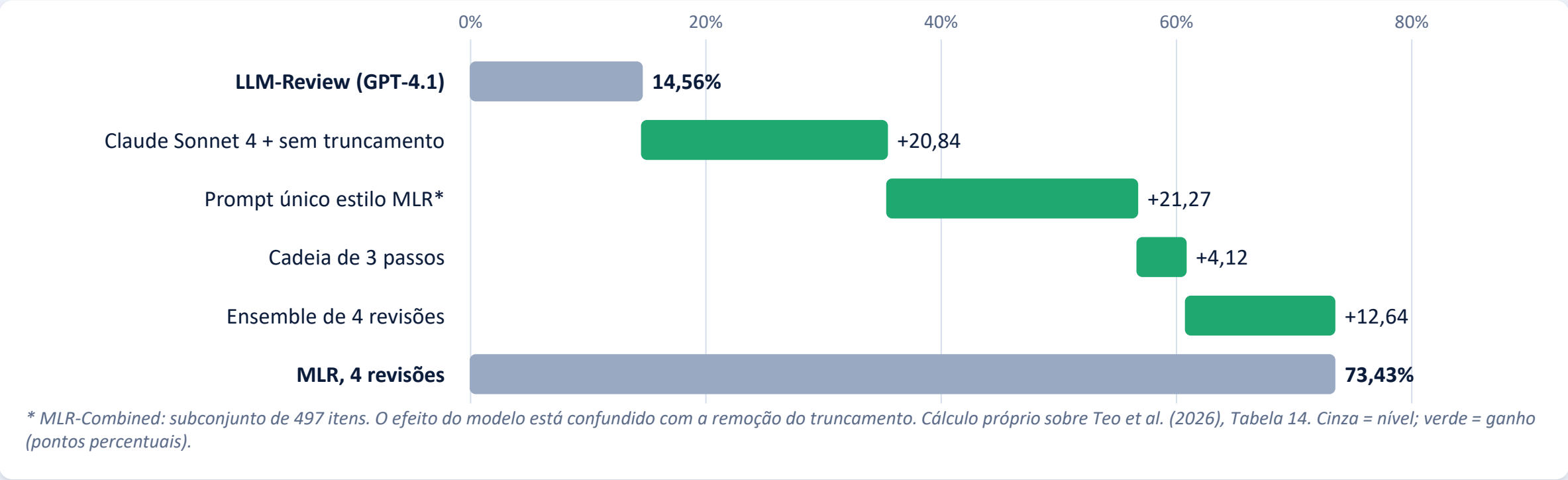
Implicações positivas

- ◆ Recall pode estar subestimado
- ◆ Ensemble ajuda mesmo com itens difíceis

Implicações negativas

- Retornos decrescentes do ensemble
- Limite superior assume sensibilidade uniforme

Dos 59 pontos ganhos no d0, a cadeia responde por 4



Por que importa

Separar modelo, truncamento, cadeia e ensemble mostra de onde vem o ganho.

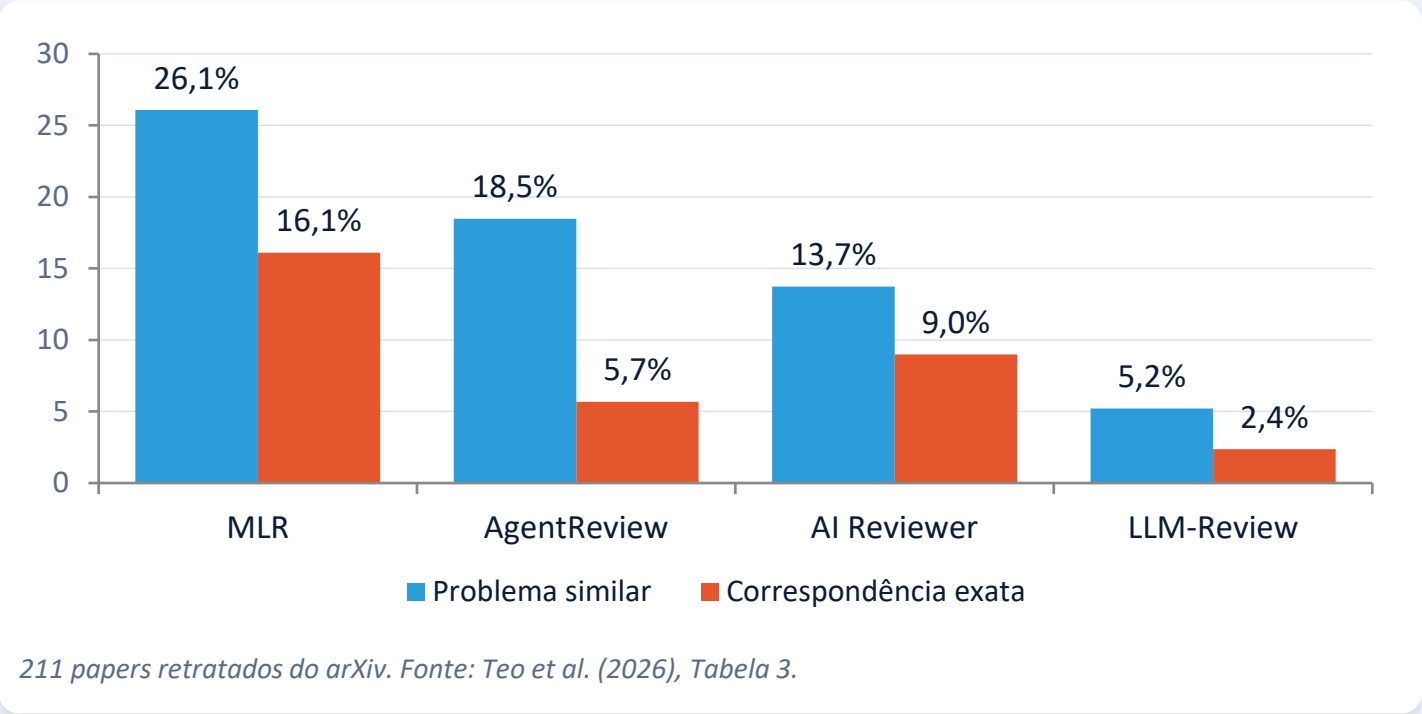
Implicações positivas

- ◆ Cadeia pesa nos erros distantes (d4: $12,85 \times 3,77$)
- ◆ Prompt único custa $\sim 1/3$

Implicações negativas

- Efeito do modelo confundido com truncamento
- MLR-Combined em subconjunto

Papers reais retratados (WithdrarXiv-Check)



Sistema	Similar (%)	Exato (%)
MLR	26,07	16,11
AgentReview	18,48	5,69
AI Reviewer	13,74	9,00
LLM-Review	5,21	2,37

Por que importa

Erros reais são o teste de fogo: a vantagem do MLR encolhe para 5–8 pontos.

Implicações positivas

- ◆ MLR ainda lidera os quatro sistemas
- ◆ Critérios 'similar' e 'exato' reportados

Implicações negativas

- 'Similar' foi definido após ver resultados
- Domínio fora de ML e limite de 10 págs.

Foco da revisão frente aos humanos

Sistema (KL; menor = mais parecido)	Pontos fortes	Fraquezas	Média
MLR	0,316	0,164	0,240
LLM-Review	0,172	0,072	0,121
AI Reviewer	0,210	0,306	0,266
AgentReview	0,299	0,058	0,178
Claude Sonnet 4.5 (prompt simples)	0,125	0,069	0,097

O MLR enfatiza validade e experimentos; humanos enfatizam clareza e novidade.

Por que importa

Mostra se a IA enxerga o que os humanos enxergam, ou algo diferente.

Implicações positivas

- ◆ Perspectiva complementar à humana
- ◆ Foco em validade e experimentos

Implicações negativas

- Divergência pode ser só menor alinhamento
- Hipóteses não separadas pelo paper

Alinhamento com as notas humanas

Pearson	MLR	LLM-Review	AI Reviewer	AgentReview	Referência humana
NeurIPS 2024	0,451	0,358	0,328	0,167	0,781
ICML 2025	0,429	0,169	0,439	0,006	0,684
ICLR 2025	0,586	-0,013	0,538	0,195	0,742

No ICML o AI Reviewer vence (Pearson 0,439; Spearman 0,416 contra 0,429 e 0,333 do MLR). BERTScore entre humanos: apenas 12–18.

Por que importa

Mostra o quanto as notas da IA acompanham o julgamento humano.

Implicações positivas

- ◆ MLR lidera em 2 de 3 venues
- ◆ Alinhamento moderado a bom (0,43–0,59)

Implicações negativas

- Abaixo da referência humana (0,68–0,78)
- Notas altas para papers rejeitados

Manipulação por injeção de texto

Sistema	Nota antes	Depois	Variação	Variação (nota própria)
MLR	7,30 ± 1,10	8,00 ± 2,88	0,70 ± 2,46	0,26 ± 1,85
LLM-Review	6,88 ± 1,32	9,30 ± 1,47	2,42 ± 1,47	n/d
AI Reviewer	7,16 ± 1,05	7,88 ± 0,62	0,72 ± 0,98	1,32 ± 1,09
AgentReview	5,51 ± 0,69	7,19 ± 0,98	1,68 ± 1,09	0,43 ± 0,69

50 papers rejeitados; texto em fonte minúscula e branca após a conclusão. MLR detectou a manipulação em 8 de 50.

Por que importa

Um revisor que lê o texto do autor tem uma superfície de ataque real.

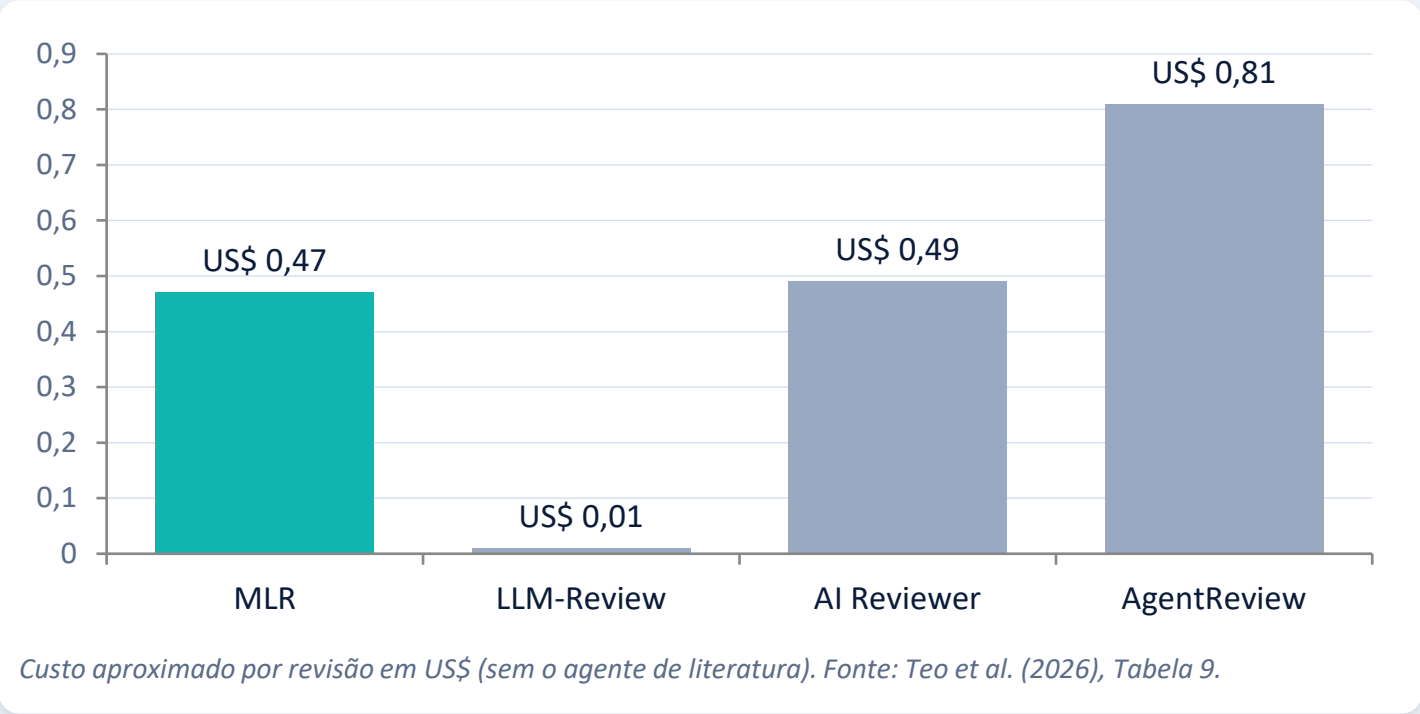
Implicações positivas

- ◆ MLR teve a menor variação média
- ◆ Entrada em imagem ajuda a detectar

Implicações negativas

- Todos os sistemas foram influenciados
- Desvio-padrão alto no MLR (2,46)

Custo comparado por revisão



Sistema	Entrada	Saída
MLR	189.062	3.913
LLM-Review	6.517	740
AI Reviewer	403.654	12.448
AgentReview	310.964	3.404

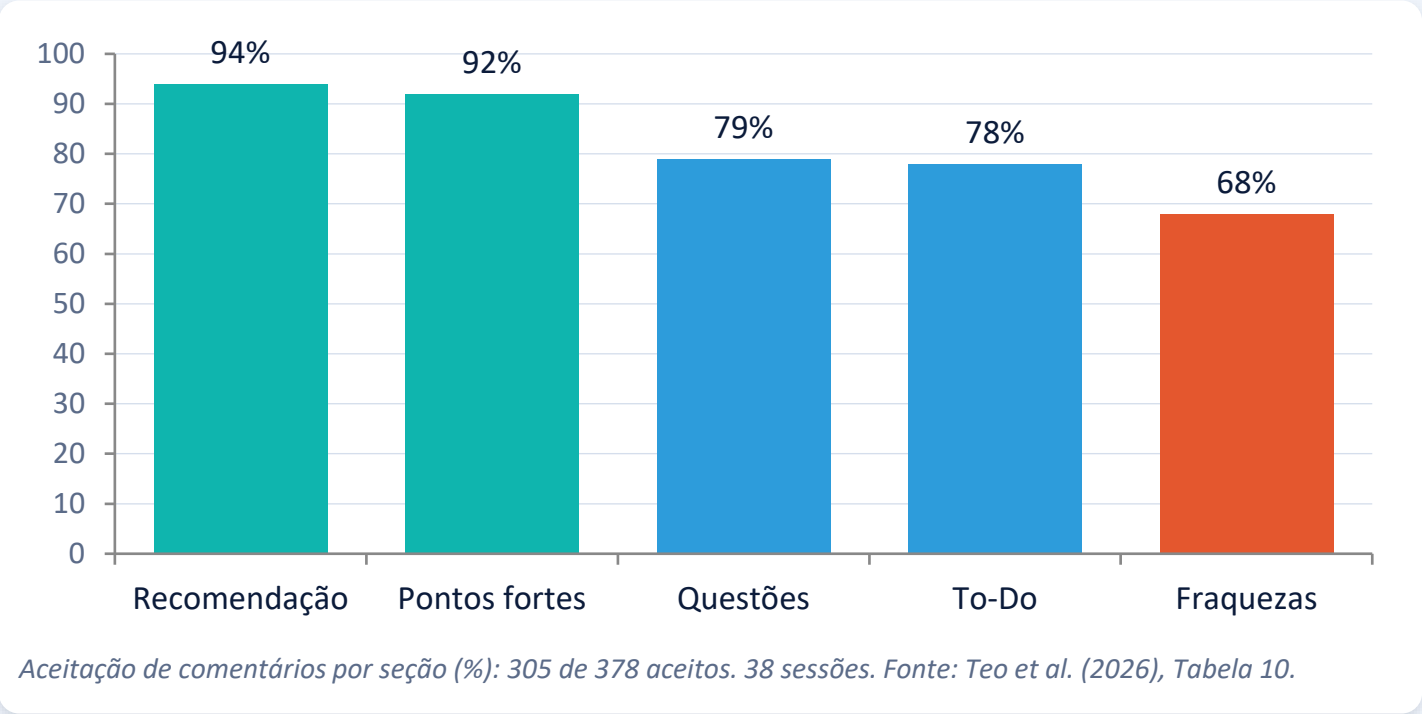
Por que importa

Custo e desempenho precisam ser lidos juntos para decidir adoção.

- Implicações positivas**
- ◆ MLR custa menos que AgentReview
 - ◆ Metade dos tokens do AI Reviewer

- Implicações negativas**
- LLM-Review é 47× mais barato
 - Preço por token do Sonnet é maior

Estudo com usuários: 81% de aceitação



Métrica (1–5)	Média
Benefício percebido	4,25
Melhoria do manuscrito	4,04
Intenção de reuso	3,74
Utilidade	3,70
Criticidade (3 = equilibrado)	2,88

Por que importa

Único teste do sistema completo de três agentes, porém qualitativo.

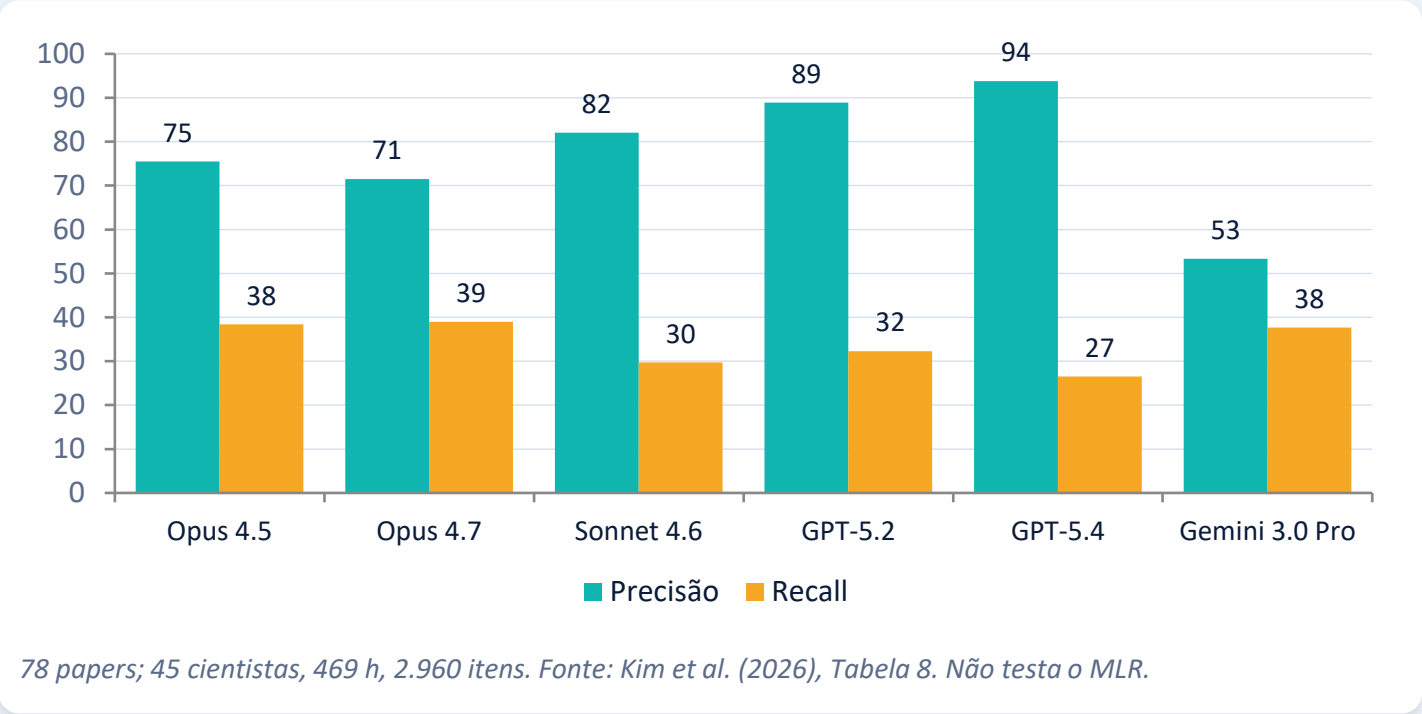
Implicações positivas

- ◆ Alta aceitação em fortes e recomendação
- ◆ Autores reportam melhoria do manuscrito

Implicações negativas

- Fraquezas, que carregam os erros: 68%
- Sem grupo controle; autosseleção

PeerReview Bench: precisão e recall reais



Revisor	P	R	F1
Claude Opus 4.5	75,49	38,39	50,89
Claude Opus 4.7	71,47	39,00	50,46
Claude Sonnet 4.6	82,03	29,74	43,65
GPT-5.2	88,92	32,28	47,37
GPT-5.4	93,81	26,55	41,38
Gemini 3.0 Pro	53,35	37,65	44,14

Por que importa

Dá a ordem de grandeza de precisão e recall que o paper da Sakana não mede.

Implicações positivas

- ◆ Revisores de IA de fronteira são úteis
- ◆ Precisão alta em alguns modelos

Implicações negativas

- Recall de 27–39% das críticas validadas
- Estudo não testa o MLR

Qualidade, sobreposição e modos de falha

Indicador	Resultado
Correção dos itens	IA 82–86% × humano top 92,3%
Itens totalmente positivos	GPT-5.2 60,0% × humano top 48,2%
Sobreposição entre revisores	IA-IA ≈ 21% × humano-humano ≈ 3%
Itens de IA sem par humano	26%, dos quais 81,8% corretos

Modos de falha (nº de comentários)

- Normas da área desalinhadas (54+46)
- Leitura errada: 'o paper não diz X' (37)
- Redundância entre revisores de IA (28)
- Comentários vagos ou verbosos (24)

Pontos fortes (nº de comentários)

- ◆ Rigor estatístico (45)
- ◆ Inspeção de código (28)
- ◆ Profundidade de domínio (27)
- ◆ Consistência interna (15)

Por que importa

Painel só de IA reduziria a diversidade de perspectivas da revisão.

Implicações positivas

- ◆ IA acha 26% de itens que humanos não acham
- ◆ Forte em estatística e código

Implicações negativas

- 14–18% de itens incorretos
- Redundância entre revisores de IA

Matriz de alegações e evidências

Alegação	Força	Ressalva
73% em alegações centrais	Alta (sintético)	Ensemble e fatia favorável; total 40,95%
"Três agentes"	Baixa	O benchmark usa só o Review Agent
Superior às baselines	Alta	Modelos e truncamento diferentes; juiz de outra família
Design contribui além do modelo	Moderada	Modelo confundido com truncamento; subconjunto
Alinha com notas humanas	Moderada	0,43–0,59 contra 0,68–0,78 da referência humana
Detecta erros reais	Baixa a moderada	16,11% (exato) e 26,07% (similar)
Mais robusto à injeção	Baixa	8 de 50; desvio-padrão alto
Beneficia pesquisadores	Sugestiva	38 sessões, sem controle; fraquezas 68%
Baixo custo	Alta	Ensemble de 4 multiplica por ~4
Precisão dos apontamentos	Nenhuma	Não medida: lacuna central

Por que importa

Cada alegação recebe uma força de evidência, não só um número.

Implicações positivas

- ◆ Separa o que se sustenta do que não
- ◆ Torna o julgamento explícito

Implicações negativas

- Classificação de força é julgamento próprio
- Pode mudar com novos dados

Pontos fortes, cautelas e a manchete

- ◆ Fortes: TMLR, preprint aberto, ablação de modelo, juiz validado, limitações discutidas
- Cautela: autoria do fornecedor, benchmark próprio, dados 'mediante solicitação'
- ▲ Cautela: precisão não medida; juiz de outra família; imprensa secundária repete a manchete
- Formulação precisa: um agente Claude Sonnet 4, 4 revisões, detectou 73,43% das contradições sintéticas na alegação principal
- ✦ Em erros reais: 16–26%; falsos alarmes não medidos

Por que importa

A credibilidade é alta para o que se mede e baixa para o que a manchete sugere.

Implicações positivas

- ◆ Resultado reproduzível em princípio
- ◆ Publicação com revisão por pares

Implicações negativas

- Autoria do fornecedor
- Dados e código só sob solicitação

Limitações reconhecidas pelos autores

- ▲ Erros sintéticos podem ser mais fáceis que erros genuínos
- Em erros reais a vantagem sobre as baselines é menos pronunciada
- ▶ Análise restrita a ML; sem comparação com sistemas proprietários
- ✦ Todos os sistemas, inclusive o MLR, são vulneráveis a manipulação
- Análise profunda só no texto principal de até 10 páginas; código e dados sob solicitação

Por que importa

Declarar limites é parte da credibilidade do resultado.

Implicações positivas

- ◆ Transparência dos autores
- ◆ Facilita replicações focadas

Implicações negativas

- Replicação independente depende dos autores
- Escopo restrito a ML

Ameaças à validade (análise própria)

Tipo	Ameaça	Efeito provável
Construto	Erro sintético ≠ real; critério 'ao menos 1 de k' favorece sensibilidade	Superestima o recall em uso real
Construto	Precisão não medida	Custo de triagem de falsos alarmes desconhecido
Interna	Modelo e truncamento confundidos; baselines com modelos diferentes	Atribuição do ganho incerta
Interna	Critério 'similar' definido após ver resultados	Possível otimismo (mitigado)
Externa	ML, 10 páginas, Sonnet 4 de 2025	Generalização limitada
Estatística	~250 itens em d0; 18 em d6; juiz com sensibilidade 86,8%	Intervalos largos; recall possivelmente subestimado

Por que importa

Mostra onde o resultado pode estar enviesado e em que direção.

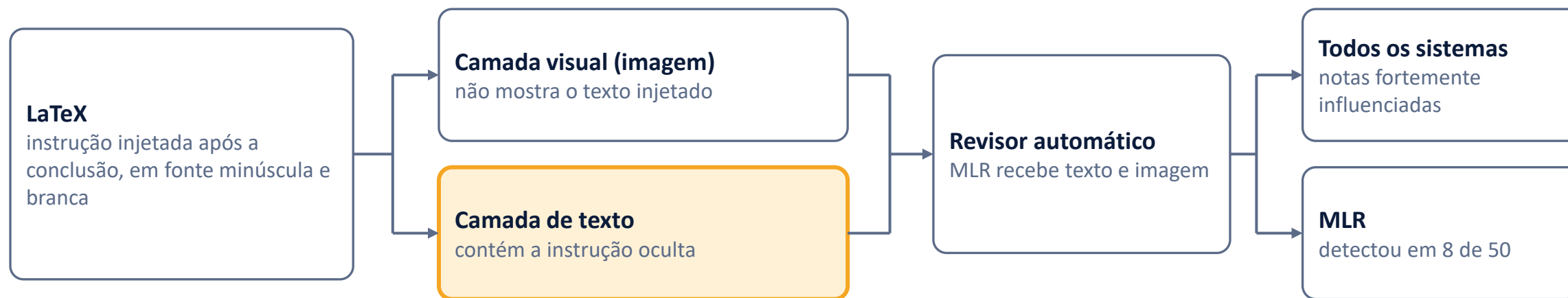
Implicações positivas

- ◆ Torna os vieses explícitos
- ◆ Aponta o que medir a seguir

Implicações negativas

- Vários vieses empurram o recall para cima
- Poucos vieses empurram para baixo

Injeção de prompt: texto oculto no PDF



Hipótese dos autores: o desenho do MLR e a entrada em texto e imagem explicam a detecção parcial.

Fonte: Teo et al. (2026), Seção 4.5 e Apêndice A.6.

Por que importa

A instrução existe na camada de texto do PDF, mas não na imagem das páginas.

Implicações positivas

- ◆ Entrada em texto e imagem ajuda a detectar
- ◆ Defesa possível: comparar as duas camadas

Implicações negativas

- 18 manuscritos reais com prompts ocultos (jul/2025)
- Detecção explícita em só 8 de 50

Lacunas: fechadas e remanescentes

Lacuna	Status	Observação
Tabela 14, ensembles, ablação da cadeia, A.6, fonte do AAAI-26	Fechada	Transcritas e conferidas no paper
Precisão do MLR	Aberta	Evidência externa só dimensiona outros revisores
Texto exato dos prompts	Parcial	Só formatos de saída do 2º e 3º passo
Figura 28 (1 a 6 revisões)	Aberta	Valores da curva não recuperados
Apêndices B.2 a B.4; A.5; A.7	Aberta	Só resultado resumido da variante
Páginas além da 10ª no WithdrarXiv	Aberta	Tratamento não esclarecido
Revisões do TMLR (OpenReview)	Aberta	Acesso bloqueado
Post original do Substack	Aberta	Não localizado

Por que importa

Dizer o que não se sabe é parte do resultado.

Implicações positivas

- ◆ Cinco lacunas fechadas com o paper
- ◆ Lista objetiva do que falta

Implicações negativas

- Precisão segue sem resposta
- Prompts completos ainda indisponíveis

Perspectivas de aprimoramento

Apontadas pelos autores

- ▶ Benchmark com erros mais naturais (retratações)
- ▶ Outros domínios além de ML
- ▶ Robustez a manipulação
- ▶ Comparar com sistemas proprietários
- ▶ Literatura com recuperação mais forte (PaperQA)
- ▶ Trocar o modelo para reduzir custo

Propostas desta análise (inferência própria)

- Medir precisão e falsos alarmes
- Reportar recall de 1 revisão com IC
- Juízes de várias famílias
- Defesas contra injeção (texto × imagem)
- Execução de código e reprodutibilidade
- Documentos longos, figuras e tabelas
- Reavaliar com modelos mais novos
- Isolar o efeito da cadeia
- Calibrar severidade por área

Por que importa

A agenda separa o que os autores propõem do que esta análise acrescenta.

Implicações positivas

- ◆ Caminho claro para fechar a lacuna de precisão
- ◆ Defesas práticas contra injeção

Implicações negativas

- Itens da análise são inferência própria
- Exigem dados e anotação humana

Uso responsável e governança

- Tratar o MLR como apoio ao revisor humano, com transparência de uso
- ▶ Supervisão humana e conformidade com as regras de cada evento
- ▲ Painéis só de IA reduzem a diversidade de perspectivas (Kim et al., 2026)
- ✦ Risco do lado do autor: refinar manuscritos contra checagens automáticas
- Risco do lado do revisor: delegar a revisão sem ler o artigo

Por que importa

Sem governança, um ganho de eficiência vira risco de integridade científica.

Implicações positivas

- ◆ Políticas de ICLR, NeurIPS, ICML e AAAI já existem
- ◆ Humano no circuito preserva a responsabilidade

Implicações negativas

- Uso não declarado e ao arrepio das regras
- Notas generosas e injeção de prompt

Requisitos de reprodução e implantação

Requisito	Detalhe
Modelos	Claude Sonnet 4 e Haiku 3.5; para o benchmark: GPT-4.1, Gemini 2.5 Pro e o3
Dados	LaTeX de papers do arXiv sob licenças CC; compilação de LaTeX
Prompts	Figuras 12, 13, 17 e 18 a 22 do apêndice do paper
Entrada	PDF com texto e imagem; texto principal até 10 páginas; ML
Custo	≈ US\$ 0,47 por revisão; ≈ US\$ 1,9 com 4 revisões; +≈ US\$ 0,97 com literatura
Disponibilidade	Dados e código mediante solicitação aos autores
Governança	Declaração de uso, supervisão humana, regras do evento

Por que importa

Quem for adotar deve medir recall e precisão com erros reais próprios.

Implicações positivas

- ◆ Sem GPU nem treinamento
- ◆ Custo previsível por artigo

Implicações negativas

- Replicação exata exige as versões do paper
- Dados e código só sob pedido

Conclusão

O MLR é sólido para achar contradições plantadas na alegação central, mas a manchete de 73% descreve o melhor caso.

● O que se sustenta

- ◆ 73,43% em d0 (4 revisões, sintético)
- ◆ Vantagem sobre as baselines
- ◆ Custo $\approx$ US\$ 0,47 por revisão

● O que não se sustenta

- 'Três agentes': foi só um
- 40,95% no total; 16–26% em erros reais
- Falsos alarmes: não medidos

● Como usar

- ▲ Segunda leitura, com humano no circuito
- ▲ Medir precisão em dados próprios
- ▲ Defender-se de injeção de prompt

Por que importa: A principal lacuna a fechar é a medição de falsos alarmes.

Fontes das notícias

Fonte primária

Teo, Yamada, Kotyan, Imajuku, Clanuwat (2026). Beyond Imitation: A Framework and Benchmark for LLM-Assisted Peer Review. TMLR (07/2026); arXiv:2610.11087 (8 Oct 2026). arxiv.org/abs/2610.11087 · openreview.net/forum?id=7iX2Z2bPFB

Cobertura secundária (a conferir)

- ▲ MarkTechPost (10 Oct 2026). Sakana AI's LLM Peer Review System Catches 73% of Core-Claim Errors. marktechpost.com
- ▲ FourWeekMBA (2026). Sakana AI Reviewer Catches 73% of Planted Main-Claim Errors. fourweekmba.com
- ▲ Beri (12 Oct 2026). Sakana's AI Reviewer Caught 73% of Planted Errors, 16% of Real Ones. beri.net
- ▲ Remio (2026). Sakana AI Peer Review Catches 73% of Core Errors, but Real Papers Expose the Limit. remio.ai
- ▲ Tech AI Magazine (2026). Sakana AI's Review System Achieves 73% Error Detection Rate. techaimag.com

O post original do Substack citado na alegação não foi localizado; a verificação se apoia no paper e nas fontes acima.

Referências bibliográficas (1/3)

- Anthropic (2024). Claude 3 model card.
- Anthropic (2025). System card: Claude Opus 4 & Claude Sonnet 4.
- Avaliação independente do AI Scientist (2025). Evaluating Sakana's AI Scientist: Bold Claims, Mixed Results, and a Promising Future? arXiv:2502.14297.
- Benchmark centrado em processo (2026). Beyond Final Decisions: A Process-Centric Benchmark for Transparent AI-Assisted Peer Review. arXiv:2609.05947.
- Beri, R. (2026). Sakana's AI Reviewer Caught 73% of Planted Errors, 16% of Real Ones. beri.net (secondary).
- Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; Wortman Vaughan, J. (2023). The NeurIPS 2021 consistency experiment. arXiv:2306.03262.
- Biswas, J. et al. (2026). AI-assisted peer review at scale: The AAI-26 AI review pilot. arXiv:2604.13940.
- Bougie, N.; Watanabe, N. (2025). Generative reviewer agents: Scalable simulacra of peer review. EMNLP 2025 Industry, 98–116.
- Chiang, C.-H.; Lee, H. (2023). Can large language models be an alternative to human evaluations? ACL 2023.
- Cortes, C.; Lawrence, N. D. (2021). Inconsistency in conference peer review: revisiting the 2014 NeurIPS experiment. arXiv:2109.09774.
- D'Arcy, M.; Hope, T.; Birnbaum, L.; Downey, D. (2024). MARG: Multi-agent review generation for scientific papers. arXiv:2401.04259.
- FourWeekMBA (2026). Sakana AI Reviewer Catches 73% of Planted Main-Claim Errors (secondary).
- Gao, X. et al. (2025). ReviewAgents: Bridging the gap between human and AI-generated paper reviews. arXiv:2503.08506.
- Gao, Z.; Brantley, K.; Joachims, T. (2024). Reviewer2: Optimizing review generation through prompt generation. arXiv:2402.10886.
- Hidden Prompts (2025). Hidden prompts in manuscripts exploit AI-assisted peer review. arXiv:2507.06185.

Referências bibliográficas (2/3)

- Jin, Y. et al. (2024). AgentReview: Exploring peer review dynamics with LLM agents. EMNLP 2024, 1208–1226.
- Keshav, S. (2007). How to read a paper.
- Kim, S. et al. (2026). On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. arXiv:2605.20668.
- Lála, J. et al. (2023). PaperQA: Retrieval-augmented generative agent for scientific research. arXiv:2312.07559.
- Liang, W. et al. (2024a). Can large language models provide useful feedback on research papers? NEJM AI 1(8), A10a2400196.
- Liu, R.; Shah, N. B. (2023). ReviewerGPT? An exploratory study on using LLMs for paper reviewing. arXiv:2306.00622.
- Lu, C. et al. (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv:2408.06292.
- Lu, C. et al. (2026). Towards end-to-end automation of AI research. Nature 651(8107), 914–919.
- MarkTechPost (2026). Sakana AI's LLM Peer Review System Catches 73% of Core-Claim Errors. 10 Oct 2026 (secondary).
- POMP (2026). Jagged AI in Scientific Peer Review: Evidence from POMP Data Analysis. arXiv:2605.07855.
- PRISM (2026). PRISM: A Multi-Dimensional Benchmark for Evaluating LLM Peer Reviewers. arXiv:2605.26730.
- R&D World (2025). Sakana AI claims first fully AI-generated paper passes peer review at top AI conference (secondary).
- Rao, D. et al. (2024). WithdrarXiv: A large-scale dataset for retraction study. arXiv:2412.03775.
- ReviewBench (2026). An Extensible Framework for Benchmarking Human and AI Manuscript Review. bioRxiv 10.64898/2026.04.17.719279.
- ReviewGrounder (2026). Improving Review Substantiveness with Rubric-Guided, Tool-Integrated Agents. arXiv:2604.14261.

Referências bibliográficas (3/3)

- Robertson, Z. (2023). GPT-4 is slightly helpful for peer-review assistance. arXiv:2307.05492.
- Russo, G. et al. (2025). The AI review lottery. Proc. ACM HCI 9(7), 1–28.
- Shin, H. et al. (2025). Mind the blind spots: A focus-level evaluation framework for LLM reviews. EMNLP 2025, 35618–35644.
- Skarlinski, M. D. et al. (2024). Language agents achieve superhuman synthesis of scientific knowledge. arXiv:2409.13740.
- Son, G. et al. (2025). SPOT: A benchmark for automated verification of scientific research. arXiv:2505.11855.
- Teo, R. S. Y.; Yamada, Y.; Kotyan, S.; Imajuku, Y.; Clanuwat, T. (2026). Beyond Imitation: A Framework and Benchmark for LLM-Assisted Peer Review. TMLR 07/2026; arXiv:2610.11087.
- Thakkar, N. et al. (2025). Can LLM feedback enhance review quality? A randomized study of 20k reviews at ICLR 2025. arXiv:2504.09737.
- Wallis, S. (2013). Binomial confidence intervals and contingency tests. J. Quantitative Linguistics 20(3), 178–208.
- Xi, S.; Rao, V.; Payan, J.; Shah, N. B. (2025). FLAWS: A benchmark for error identification and localization in scientific papers. arXiv:2511.21843.
- Ye, R. et al. (2024). Are we there yet? Revealing the risks of utilizing LLMs in scholarly peer review. arXiv:2412.01708.
- Yuan, W.; Liu, P.; Neubig, G. (2022). Can we automate scientific reviewing? JAIR 75, 171–212.
- Zhang, T. M.; Abernethy, N. F. (2025). Reviewing scientific papers for critical problems with reasoning LLMs. arXiv:2505.23824.
- Zhu, M.; Weng, Y.; Yang, L.; Zhang, Y. (2025). DeepReview: Improving LLM-based paper review with human-like deep thinking process. ACL 2025, 29330–29355.

Glossário (1/2)



MLR

Multi-Layered Review: sistema de revisão em camadas da Sakana AI



Distância (d)

Menor caminho no grafo até uma alegação principal; d0 é a central



Contradição sintética

Erro lógico inserido de forma controlada por um LLM



Ensemble de k revisões

k revisões independentes; conta se qualquer uma acertar



Recall

Fração dos erros conhecidos que o revisor encontra



Precisão

Fração dos apontamentos do revisor que são corretos



LLM-juiz

Modelo que decide se a revisão detectou o erro



KL (Kullback-Leibler)

Divergência entre distribuições; compara o foco dos revisores



Injeção de prompt

Instrução oculta no documento que tenta manipular o revisor

Glossário (2/2)

WithdrarXiv-Check

Papers retirados do arXiv com comentário de retratação

PeerReview Bench

Benchmark validado por cientistas, com precisão e recall

Three-Pass Approach

Método de leitura em três passes de Keshav (2007)

Grafo de conhecimento

Rede de alegações, evidências, métodos e implementações do paper

Appendix Agent

Agente (Haiku 3.5) que resume o apêndice do artigo

Review Agent

Agente (Sonnet 4) que executa a cadeia de 3 passos

Truncamento

Corte do texto de entrada por limite de contexto

IC de Wilson

Intervalo de confiança para proporções em amostras pequenas

Retratação

Retirada formal de um artigo por erro ou má conduta

Nota de verificação e direitos autorais

Fontes primárias lidas

Texto principal e apêndices A.1 a A.4 do paper, trechos de A.5, A.6 e B.1 (Teo et al., 2026) e o texto principal de Kim et al. (2026).

Fontes secundárias

MarkTechPost, FourWeekMBA e Beri: usadas só para conferência cruzada.

Não verificados

Substack original, revisões do TMLR no OpenReview, Figura 28 e Apêndices A.5, A.7, B.2 a B.4 em detalhe.

Cálculos próprios

Sinalizados nos slides como 'cálculo próprio' ou 'inferência própria'.

© Aridio Silva · @aridiosilva · 13-Outubro-2026

Aridio Silva, Pesquisador Independente · ORCID 0009-0008-2411-6995 · Texto, tabelas, gráficos, diagramas e imagens deste infográfico são de autoria de Aridio Silva.