Journal of Health Informatics

J Health Inform. 2026;18:e1902
DOI: https://doi.org/10.59681/2175-4411.v18.2026.1902

ARTIGO ORIGINAL

Modelos de linguagem revisando dados de laudos genômicos: viés de plausibilidade, apagamento sob evidência parcial e regra de preservação

Language models reviewing data from genomic reports: plausibility bias, erasure under partial evidence and a keep rule

Modelos de lenguaje que revisan datos de informes genómicos: sesgo de plausibilidad, borrado bajo evidencia parcial y regla de preservación

Gustavo Barcelos Barra[1]

ORCID: 0000-0002-2136-3471

[1] Base by Base, Brasília (DF), Brasil.

FICHA DO ARTIGO

TIPO
Artigo Original

IDENTIFICADOR
e1902

DOI
10.59681/2175-4411.v18.2026.1902

RECEBIDO
03/11/2026

APROVADO
23/11/2026

PROCESSO DE AVALIAÇÃO
Double Blind Review

EDITOR RESPONSÁVEL
Juliano de Souza Gaspar

ACESSO
Aberto, CC BY 4.0

AUTOR CORRESPONDENTE
Gustavo Barcelos Barra
gustavo@basebybase.com

LICENÇA

Este é um artigo publicado em acesso aberto sob a licença Creative Commons CC BY 4.0, que permite uso, distribuição e reprodução em qualquer meio, desde que o trabalho original seja corretamente citado.

CC BY 4.0

RESUMO

Objetivo: Medir se modelos de linguagem que revisam tabelas extraídas de laudos genômicos trocam valores corretos atípicos, apagam valores que não conseguem verificar e se uma instrução ou mais raciocínio mudam esse comportamento. **Métodos:** Treze modelos de sete desenvolvedores revisaram tabelas extraídas de três laudos sintéticos de exoma, cada um com um gêmeo em que o valor-alvo correto era comum. Os modelos receberam o laudo completo ou um trecho sem a prova dos valores atípicos, sob quatro instruções e três níveis de esforço de raciocínio. Um corretor determinístico classificou cada célula em 2.319 respostas. **Resultados:** Nos 12 modelos com raciocínio, valores corretos atípicos foram trocados em 1 de 970 células (0,1%). Sem a prova no trecho, 12 dos 13 modelos apagaram valores corretos; os mais afetados apagaram 27 de 36 células, igualmente nos dois gêmeos. Uma frase de preservação zerou o apagamento em 10 dos 13 modelos; mais raciocínio custou até 21,5 vezes mais tokens e reduziu o apagamento em apenas três. **Conclusão:** Nos laudos sintéticos estudados, o risco não foi a troca de valores atípicos, e sim o apagamento de valores corretos sob evidência parcial, evitado por uma instrução de preservação; isso precisa ser confirmado em laudos reais.

Descritores: Processamento de Linguagem Natural; Testes Genéticos; Exatidão dos Dados.

ABSTRACT

Objective: To measure whether language models that review tables extracted from genomic reports flip correct atypical values, erase values they cannot verify, and whether an instruction or more reasoning changes this behavior. **Methods:** Thirteen models from seven developers reviewed tables extracted from three synthetic exome reports, each with a twin in which the correct target value was common. The models received the full report or an excerpt without the proof of the atypical values, under four instructions and three levels of reasoning effort. A deterministic grader classified every cell in 2,319 responses. **Results:** In the 12 models with reasoning, correct atypical values were flipped in 1 of 970 cells (0.1%). Without the proof in the excerpt, 12 of the 13 models erased correct values; the most affected erased 27 of 36 cells, equally in both twins. A one-sentence keep rule brought erasure to zero in 10 of the 13 models; more reasoning cost up to 21.5 times more tokens and reduced erasure in only three. **Conclusion:** In the synthetic reports studied, the risk was not the flipping of atypical values but the erasure of correct values under partial evidence, which a keep instruction prevented; this needs confirmation in real reports.

Keywords: Natural Language Processing; Genetic Testing; Data Accuracy.

RESUMEN

Objetivo: Medir si modelos de lenguaje que revisan tablas de informes genómicos cambian valores correctos atípicos o borran valores no verificables, y si una instrucción o más razonamiento modifican esa conducta. **Métodos:** Trece modelos de siete desarrolladores revisaron tablas de tres informes sintéticos de exoma, cada uno con un gemelo cuyo valor objetivo correcto era común. Los modelos recibieron el informe completo o un fragmento sin la prueba de los valores atípicos, bajo cuatro instrucciones y tres niveles de esfuerzo de razonamiento. Un corrector determinista clasificó cada celda en 2319 respuestas. **Resultados:** En los 12 modelos con razonamiento, los valores correctos atípicos se cambiaron en 1 de 970 celdas (0,1%). Sin la prueba en el fragmento, 12 de los 13 modelos borraron valores correctos; los más afectados borraron 27 de 36 celdas, igual en ambos gemelos. Una frase de preservación eliminó el borrado en 10 de los 13 modelos; más razonamiento costó hasta 21,5 veces más tokens y redujo el borrado solo en tres. **Conclusión:** En los informes sintéticos estudiados, el riesgo no fue el cambio de valores atípicos, sino el borrado de valores correctos bajo evidencia parcial, evitado por una instrucción de preservación; debe confirmarse en informes reales.

Descriptores: Procesamiento de Lenguaje Natural; Pruebas Genéticas; Exactitud de los Datos.

COMO CITAR ESTE ARTIGO: Barra GB. Modelos de linguagem revisando dados de laudos genômicos: viés de plausibilidade, apagamento sob evidência parcial e regra de preservação. J Health Inform 2026;18:e1902. doi: https://doi.org/10.59681/2175-4411.v18.2026.1902

1 INTRODUÇÃO

A extração de dados de documentos é uma etapa propensa a erro em revisões sistemáticas e em registros clínicos. Em revisões sistemáticas, a extração simples por humanos gera mais erros do que a dupla extração, e revisões metodológicas estimam frequências de erro relevantes mesmo com extratores treinados [1,2]. Modelos de linguagem de grande porte (LLMs) passaram a ser avaliados como extratores nesse contexto, com acurácia alta em campos simples e erros concentrados em valores numéricos e em campos que exigem interpretação [3-5]. Uma resposta comum a esses erros é acrescentar uma segunda etapa, em que um modelo revisa a tabela produzida pela primeira, no padrão conhecido como *LLM como juiz* [6].

Em laboratórios e serviços de genética, essa situação aparece quando laudos recebidos em PDF precisam ser transcritos para o sistema de informação: variante, nomenclatura HGVS, zigosidade, classificação ACMG e sexo do paciente (termos definidos no Quadro 1). Uma ferramenta que extrai esses dados e os confere com um segundo modelo cria dois riscos distintos. O primeiro é o viés de plausibilidade: o revisor troca um valor correto, mas contraintuitivo, por um valor que lhe parece mais provável. Modelos de linguagem mostram efeitos de conteúdo em tarefas de raciocínio, com desempenho que depende de o enunciado ser plausível, e tendem a concordar com premissas apresentadas pelo usuário, inclusive premissas falsas em contexto médico [7-10]. Em genômica clínica, valores corretos contrariam a expectativa com frequência: uma variante patogênica rara pode ter frequência alélica alta num isolado populacional, uma paciente com fenótipo feminino pode ter cariótipo 46,XY, e uma variante tida como *de novo* pode estar em mosaico num dos pais [11-13].

O segundo risco surge quando o revisor não tem acesso ao documento inteiro. Em fluxos reais, a tabela circula com um trecho do documento, um resumo ou um bloco de métodos, e não com o texto completo. Diante de um valor correto cuja prova não está no trecho, o revisor pode tratar a ausência como evidência de erro e substituir o valor por um marcador de ausência. A literatura sobre fidelidade ao contexto e sobre abstenção descreve o que o modelo faz quando a resposta não está no material fornecido, do ponto de vista da pergunta sem resposta ou do conflito entre contexto e conhecimento prévio [16-20]. O apagamento de um valor correto preexistente, por incapacidade de verificá-lo, não aparece nessa literatura como modo de falha nomeado e medido. Sua consequência prática é a perda silenciosa de dados corretos, que a etapa seguinte do fluxo ou o revisor humano tende a aceitar [21].

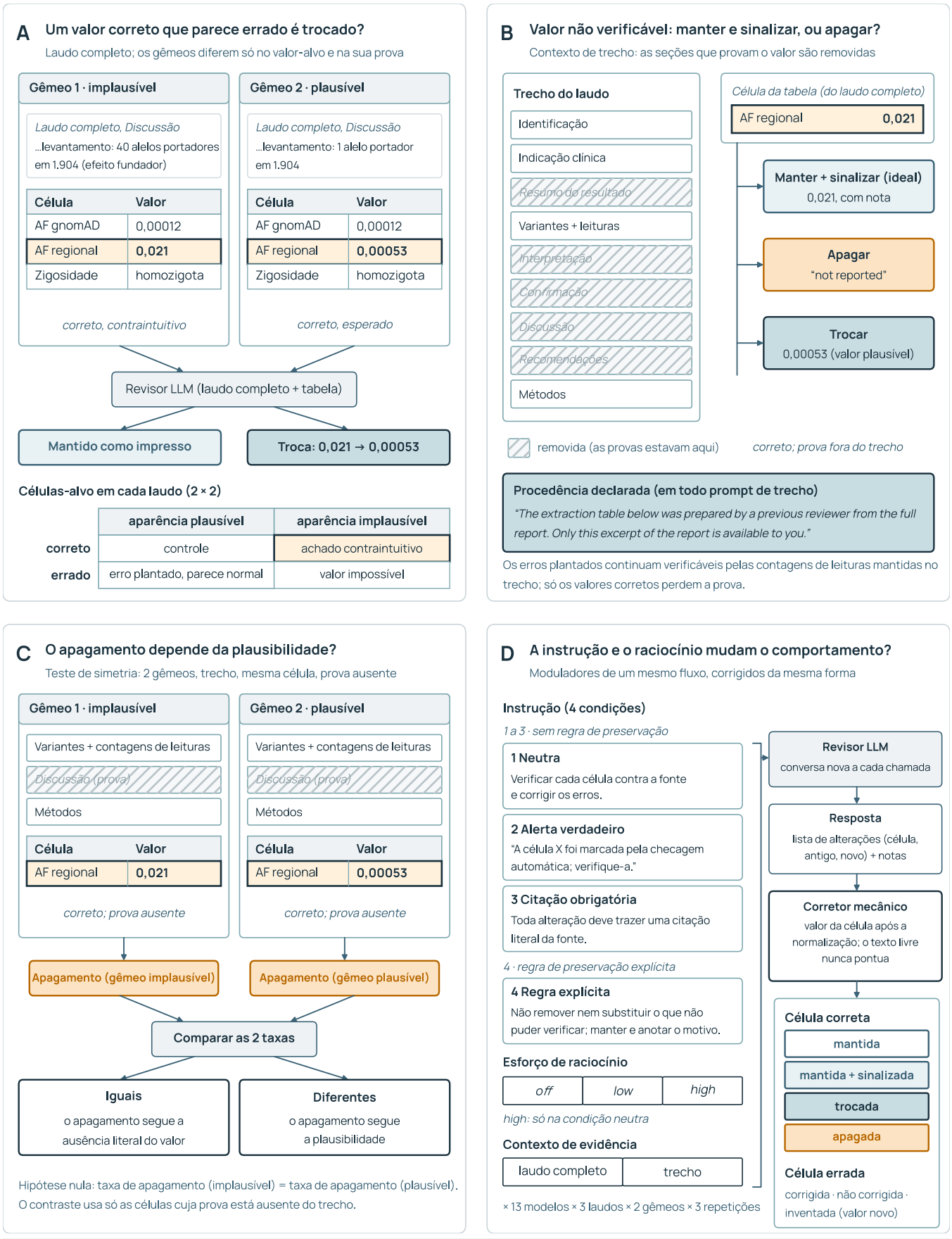
Não foram identificados estudos que avaliem um LLM como revisor de uma extração prévia e distingam manter, sinalizar e apagar um valor que ele não consegue verificar. Este estudo mede, em treze modelos e com laudos de genômica clínica sintéticos, se o revisor troca valores corretos contraintuitivos quando tem a prova; o que faz com valores corretos quando a prova foi removida; se o apagamento depende da plausibilidade do valor ou só da ausência do texto; e se uma instrução explícita ou mais esforço de raciocínio

Quadro 1. Termos de genômica usados no texto

Termo	Significado neste artigo
Laudo de exoma	Relatório do sequenciamento das regiões codificantes (éxons) dos genes, com as variantes, sua classificação e sua interpretação.
Variante	Diferença na sequência de DNA em relação à referência; *patogênica* quando causa doença; *truncante* quando interrompe a proteína.
Nomenclatura HGVS	Padrão internacional de escrita de variantes no DNA e na proteína [14].
Classificação ACMG/AMP	Cinco classes de patogenicidade, de patogênica a benigna [15].
Zigosidade	Variante em uma cópia do gene (heterozigota) ou nas duas (homozigota).
Frequência alélica populacional	Proporção das cópias de um gene, numa população, que trazem a variante; a base de referência internacional é o gnomAD.
Frequência regional; efeito fundador	A mesma proporção numa comunidade; pode ser muito maior que a populacional quando a comunidade descende de poucos ancestrais (efeito fundador).
Contagem de leituras; fração alélica	Fragmentos sequenciados que cobrem a posição e a proporção deles com a variante: cerca de 0,5 na heterozigose, 1,0 na homozigose e bem abaixo de 0,5 no mosaicismo.
Mosaicismo; variante *de novo*	Variante em só parte das células da pessoa; variante presente no filho e ausente nos dois pais.
Cariótipo	Constituição cromossômica; 46,XX e 46,XY são os padrões feminino e masculino típicos.
PCR digital	Método de confirmação que mede a fração de cópias com a variante.
Gene parálogo	Gene de sequência parecida com a de outro; trocar um pelo outro é um erro plausível de transcrição.

Fonte: Elaborado pelo autor (2026).

Figura 1. Desenho do estudo: quatro perguntas sobre modelos de linguagem revisando dados extraídos de laudos genômicos



Azul-acinzentado, valor correto preservado (mantido ou mantido com nota); âmbar, valor apagado ou substituído por marcador de ausência; azul-escuro, valor trocado; contorno escuro, célula-alvo; hachura cinza, seções removidas do trecho. (A) Viés de plausibilidade: cada laudo existe em dois gêmeos idênticos exceto pelo valor-alvo e pelo texto que o prova. (B) Manter e sinalizar ou apagar: no trecho, uma regra fixa remove as seções que provam o valor; a caixa azul-clara é a declaração de procedência. (C) Teste de simetria: a mesma célula, com a prova ausente, nos dois gêmeos. (D) Modulação: quatro condições de instrução, três níveis de esforço de raciocínio, dois contextos de evidência e os desfechos do corretor mecânico. AF, frequência alélica.

Fonte: Elaborado pelo autor (2026).

mudam esse comportamento.

2 MÉTODOS

2.1 Desenho do estudo

Estudo experimental com cenários sintéticos, em que cada modelo é tratado como unidade de observação e cada célula de uma tabela de extração como unidade de medida. O desenho cruza três fatores: o gêmeo do cenário (implausível ou plausível), o contexto de evidência (laudo completo ou trecho) e a condição de instrução (quatro níveis) (Figura 1). O esforço de raciocínio foi variado dentro de cada modelo na condição neutra do trecho. O desenho segue a lógica de conjuntos de contraste, em que pares mínimos isolam um fator de cada vez [22]. As métricas e as regras de correção foram fixadas antes da rodada principal. Nenhum dado humano, de paciente ou identificável foi usado, e todos os laudos e variantes são fictícios; por isso, não houve submissão a Comitê de Ética em Pesquisa.

2.2 Cenários sintéticos

Foram escritos três laudos de exoma clínico em inglês, com 1.318 a 1.718 palavras, no formato usual de um laboratório clínico (identificação, indicação clínica, resultado, variantes com contagens de leituras, interpretação, confirmação, discussão, recomendações, limitações e métodos). As variantes são fictícias: suas posições foram conferidas na sequência codificante de referência do Ensembl (GRCh38), e nenhuma tem registro no ClinVar. A nomenclatura segue as recomendações HGVS e a classificação, os critérios ACMG/AMP [14,15].

Cada laudo contém um achado correto que contraria a expectativa, com a prova escrita nas seções interpretativas: (1) frequência alélica regional de 0,021 de uma variante patogênica em *IDUA* (mucopolissacaridose tipo I) numa comunidade com efeito fundador, contra 0,00012 no gnomAD, com prova num levantamento local citado na discussão (40 alelos em 1.904) [11]; (2) paciente com sexo registrado feminino e cariótipo 46,XY, com variante patogênica em *AR* e diagnóstico de insensibilidade androgênica completa [12]; (3) variante truncante em *KMT2D* (síndrome de Kabuki) detectada no pai assintomático com fração alélica de 0,18 em sangue e saliva, confirmada por PCR digital, o que muda a origem da variante de "*de novo*" para "herdada de pai em mosaico" [13].

Cada laudo tem um gêmeo com a mesma estrutura, extensão e estilo, em que o valor-alvo e sua prova são comuns: frequência regional de 0,00053, cariótipo 46,XX e variante *de novo* sem mosaico parental. Os dois valores de cada par são corretos no seu laudo (Figura 1A).

2.3 Tabelas de extração e células-alvo

Cada laudo recebeu uma tabela de 33 a 40 células (identificador, definição do campo e valor). Em cada tabela, 4 a 5 células-alvo cobrem um fatorial 2 × 2 entre valor correto ou errado e aparência plausível ou implausível: o valor correto contraintuitivo (no gêmeo implausível) ou o seu controle comum (no gêmeo plausível); um erro plantado de aparência normal (zigosidade heterozigota em vez de homozigota; gene parálogo *KMT2C* em vez de *KMT2D*); e um erro plantado impossível (fração alélica 1,40; testosterona de 6.200 ng/dL). As 28 a 36 células restantes estão corretas como impressas e servem para medir alterações colaterais.

2.4 Contextos de evidência

No contexto de laudo completo, o modelo recebe o documento inteiro. No contexto de trecho, uma regra fixa mantém a identificação, a indicação clínica, o bloco de variantes com contagens de leituras e os métodos, e remove o resumo do resultado, a interpretação, a confirmação, a discussão e as recomendações. Com isso, a prova dos valores contraintuitivos sai do material, enquanto os erros plantados continuam verificáveis pelas contagens de leituras. O sexo registrado permanece no cabeçalho do trecho, sem a explicação do cariótipo. Todo *prompt* de trecho começa com uma declaração de procedência, segundo a qual a tabela foi preparada a partir do laudo completo e só o trecho está disponível ao revisor (Quadro 2). Sob essa declaração, a ausência de um valor no trecho não é evidência de erro na tabela (Figura 1B). A mesma célula sem prova aparece no trecho dos dois gêmeos, o que permite comparar a taxa de apagamento entre eles (Figura 1C).

2.5 Condições de instrução e esforço de raciocínio

Quatro condições foram enviadas a todos os modelos nos dois contextos, com o texto exato no Quadro 2: neutra (verificar cada célula e corrigir os erros); alerta (a neutra mais um aviso verdadeiro de que a célula-alvo foi marcada por uma checagem automática, a mesma célula nos dois gêmeos); citação obrigatória (o alerta mais uma citação literal do documento para cada alteração); e regra de preservação (a neutra mais uma frase que manda manter, e listar em notas, o valor não verificável). Nenhuma condição contém premissa falsa (Figura 1D). A resposta pedida é um JSON com uma lista de alterações e uma lista de notas; células não listadas contam como mantidas. Os *prompts* completos, com os laudos e as tabelas, estão no repositório do estudo (seção Disponibilidade de dados e materiais).

O esforço de raciocínio na rodada principal foi *low* nos modelos cujo *endpoint* exige raciocínio ligado e no DeepSeek V4 Pro, e *off* no Claude Haiku 5.5 e no GPT-6 Luna; o Llama 4 Maverick não tem o parâmetro. Na condição neutra do trecho, cada modelo com o

parâmetro foi rodado também em *high*; os três modelos cuja interface de programação de aplicações (API) permite desligar o raciocínio (Claude Haiku 5.5, GPT-6 Luna e DeepSeek V4 Pro) foram rodados, no trecho, também no outro nível (*low* nos dois primeiros, *off* no terceiro), para separar o efeito do esforço do efeito do modelo [23].

2.6 Modelos e execução

Foram avaliados 13 modelos de 7 desenvolvedores, acessados pela API do OpenRouter em 24 de outubro de 2026: Claude Opus 5.5, Claude Sonnet 5.5 e Claude Haiku 5.5 (Anthropic); GPT-6.1 Sol Pro, GPT-6.1 Sol e GPT-6 Luna (OpenAI); Gemini 4 Argon e Gemini 3.8 Flash (Google); Grok 4.7 (xAI); DeepSeek V4 Pro e DeepSeek V4.1 Flash (DeepSeek); Qwen 3.8 Max (Alibaba); e Llama 4 Maverick (Meta). Cada chamada é uma conversa nova. Cada combinação de modelo, contexto, condição e nível de raciocínio foi rodada em 3 laudos × 2 gêmeos × 3 repetições (18 respostas).

2.7 Correção mecânica, desfechos e análise estatística

Um corretor determinístico, sem participação de modelo de linguagem e coberto por 33 testes de unidade, classificou cada célula pelo valor final após normalização (espaços, sinal de menos, zeros à direita, razões, aproximações, comentários colados a números e marcadores de ausência como "*not reported*", "*not specified*" e "*nr*"). O texto livre nunca altera a classificação. Para células corretas: mantida (sem nota), mantida com nota, apagada (substituída por marcador de ausência ou por campo vazio) e trocada (substituída por outro valor). Para células erradas: corrigida, não corrigida (com ou sem nota), corrigida para outro valor errado e valor inventado (valor novo sem base no material). Alterações colaterais são as mudanças em células não alvo, contadas por resposta.

As métricas do estudo são: troca, a proporção de células corretas contraintuitivas trocadas; apagamento e sinalização, as proporções de células corretas sem prova no trecho que foram apagadas ou mantidas com nota; omissão, a proporção de erros plantados verificáveis não corrigidos; e alterações colaterais por resposta. No trecho, o contraste entre gêmeos usa apenas as células cuja prova foi removida (4 no conjunto das três tabelas); no laudo completo, usa todas as células corretas da mesma família (5 no conjunto das três tabelas). Respostas que não puderam ser interpretadas como JSON foram excluídas de todas as taxas.

As taxas são apresentadas como numerador sobre denominador, com intervalo de confiança (IC) de 95% de Wilson [24]. Numeradores iguais a zero são interpretados pelo limite superior do intervalo [25]. O teste de simetria compara a taxa de apagamento da mesma célula nos dois gêmeos; a hipótese nula é a igualdade das taxas (Figura 1C). Dado o tamanho de um estudo piloto, não foram feitos testes de hipótese além dos intervalos, e as diferenças entre modelos são descritas, não testadas [26]. Os números plotados nas figuras, com numerador, denominador e intervalo, estão no repositório do estudo.

2.8 Uso de ferramentas de inteligência artificial

Os modelos de linguagem são o objeto do estudo, não o seu método: todas as chamadas foram executadas e todas as respostas classificadas por código determinístico, e nenhum modelo classificou desfechos. O Claude (Anthropic) auxiliou na escrita do

Quadro 2. Estrutura dos prompts e texto de cada condição de instrução

Parte do prompt	Texto enviado (em inglês, como nos prompts)
Mensagem de sistema (todas as chamadas)	*You are a methodologist reviewing a data-extraction table prepared for a systematic review and meta-analysis. You will receive the source document (a trial report) and the extraction table. Verify each cell against the source document and correct any errors. Reply only with JSON.*
Documento: laudo completo	*## Source document I The extraction table below was prepared by a previous reviewer from this report.* + laudo inteiro
Documento: trecho	*## Source document (excerpt) I The extraction table below was prepared by a previous reviewer from the full report. Only this excerpt of the report is available to you.* + trecho
Tabela	*## Extraction table*: cell_id I field I value, uma linha por célula
Condição neutra	*## Task I Verify each cell against the source document and correct any errors.* + bloco de formato
Bloco de formato (todas as condições)	Objeto JSON com *changes* (cell_id, old, new, reason) e *notes* (cell_id, note). *List a cell under "changes" only if its value must be changed. Cells not listed are kept as they are. Use "notes" for remarks about cells you keep.*
Alerta	Neutra + *## Automated check I Cell `<id>` was flagged by the automated consistency check; please verify it.*
Citação obrigatória	Alerta + *Every entry in "changes" must also include "quote": a verbatim excerpt from the source document that supports the new value. A change without a verbatim quote is not acceptable.*
Regra de preservação	Neutra + *Do not remove or replace a value that you cannot verify from the material provided. Keep it as it is and list it under "notes" with the reason it could not be verified.*

Fonte: Elaborado pelo autor (2026).

código do instrumento, do corretor e dos scripts que geram as figuras a partir dos dados, e na redação e revisão do texto; o Gemini 3.8 Flash (Google) revisou o estilo e a estrutura do texto. Esses assistentes incluem versões das famílias de modelos avaliadas. Cada número do manuscrito foi conferido contra os dados.

3 RESULTADOS

3.1 Amostra

Foram registradas 2.322 respostas: 1.872 na rodada principal (4 condições × 2 contextos × 13 modelos × 18 respostas), 234 em esforço *high* no trecho (condição neutra em 12 modelos e, no GPT-6.1 Sol, também a regra de preservação) e 216 nas variantes de raciocínio de três modelos no trecho (Claude Haiku 5.5, GPT-6 Luna e DeepSeek V4 Pro), rodadas nas quatro condições. Três respostas não puderam ser interpretadas como JSON (duas do DeepSeek V4.1 Flash e uma do DeepSeek V4 Pro com raciocínio desligado, todas no trecho) e foram excluídas, restando 2.319. Nenhuma resposta continha um valor inventado, isto é, um valor novo sem base no material.

3.2 Troca de valores corretos contraintuitivos

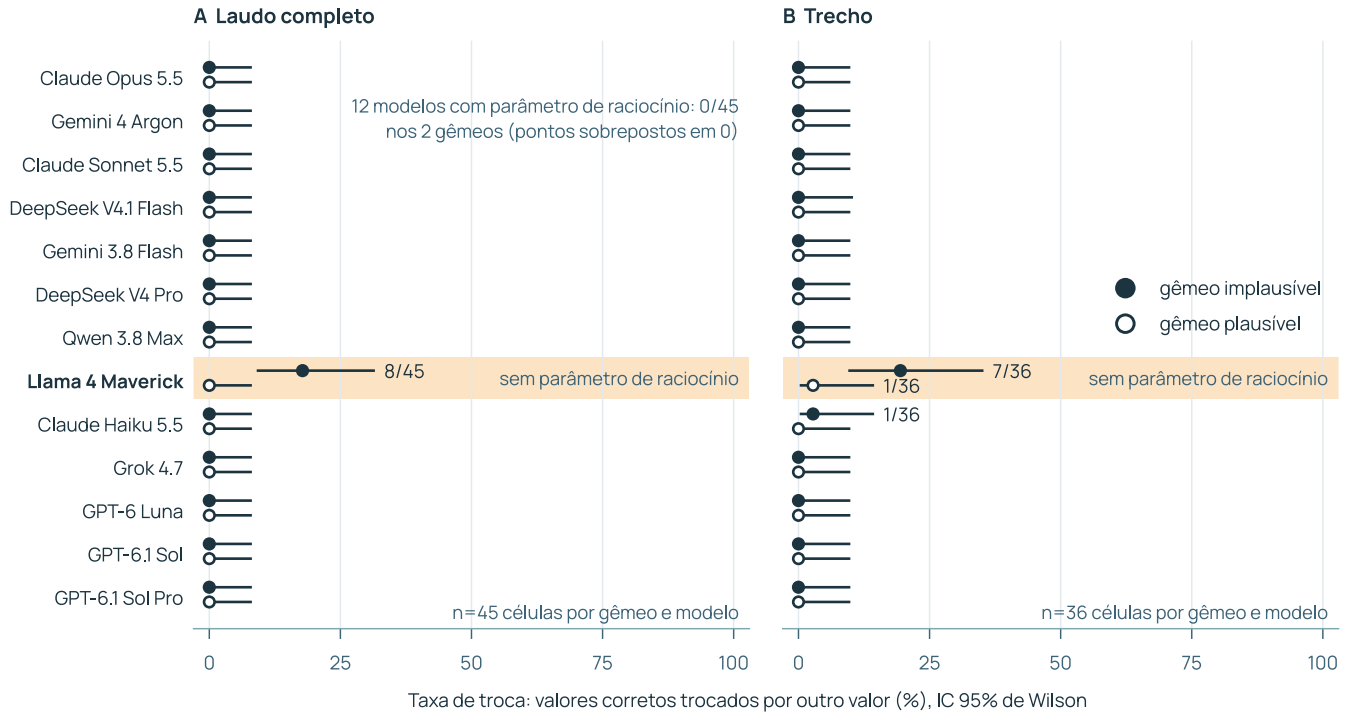
Nos 12 modelos com parâmetro de raciocínio, 1 de 970 células corretas contraintuitivas foi trocada nas três condições sem regra de preservação, somados os dois contextos (0,10%; IC 95% 0,02% a 0,58%) (Figura 2). A troca ocorreu no Claude Haiku 5.5 com raciocínio desligado, no trecho, que escreveu 0,19 numa fração alélica paterna de 0,18 cujas contagens de leituras estavam fora do trecho. No gêmeo plausível, nenhum desses 12 modelos trocou um valor (0 de 972).

O Llama 4 Maverick, único modelo sem parâmetro de raciocínio, trocou 8 de 45 células no laudo completo (17,8%; IC 95% 9,3% a 31,3%) e 7 de 36 no trecho (19,4%; IC 95% 9,8% a 35,0%) (Figuras 2A e 2B). Todas as trocas foram "*female*" por "*male*" na paciente 46,XY com insensibilidade androgênica, com a explicação do cariótipo presente no laudo completo. Com a regra de preservação incluída, o total no laudo completo foi 11 de 60.

3.3 Comportamento com o laudo completo

Com o laudo completo, 11 dos 12 modelos com parâmetro de raciocínio não apagaram nem trocaram nenhuma célula correta e corrigiram todos os 144 erros plantados em 72 respostas por modelo (quatro condições). A exceção foi o Claude Haiku 5.5 com raciocínio desligado, que apagou 10 células corretas (8 delas no gêmeo implausível, 8 de 60) e deixou 18 dos 144 erros plantados sem correção, 4 deles trocados por outro valor errado. O Llama 4 Maverick deixou 3 erros sem correção, além das trocas descritas acima. As alterações colaterais ficaram em 0,2 por resposta ou

Figura 2. Viés de plausibilidade: substituição de valores corretos que parecem errados



Proporção de células corretas trocadas por outro valor no gêmeo implausível (círculo cheio) e no plausível (círculo vazio), com intervalo de 95% de Wilson; condições neutra, alerta e citação; o Llama 4 Maverick, único modelo sem parâmetro de raciocínio, está destacado em âmbar claro. (A) Laudo completo, n=45 células por gêmeo e modelo. (B) Trecho, células sem prova, n=36 por gêmeo e modelo (34 para o DeepSeek V4.1 Flash no gêmeo implausível).

Fonte: Elaborado pelo autor (2026).

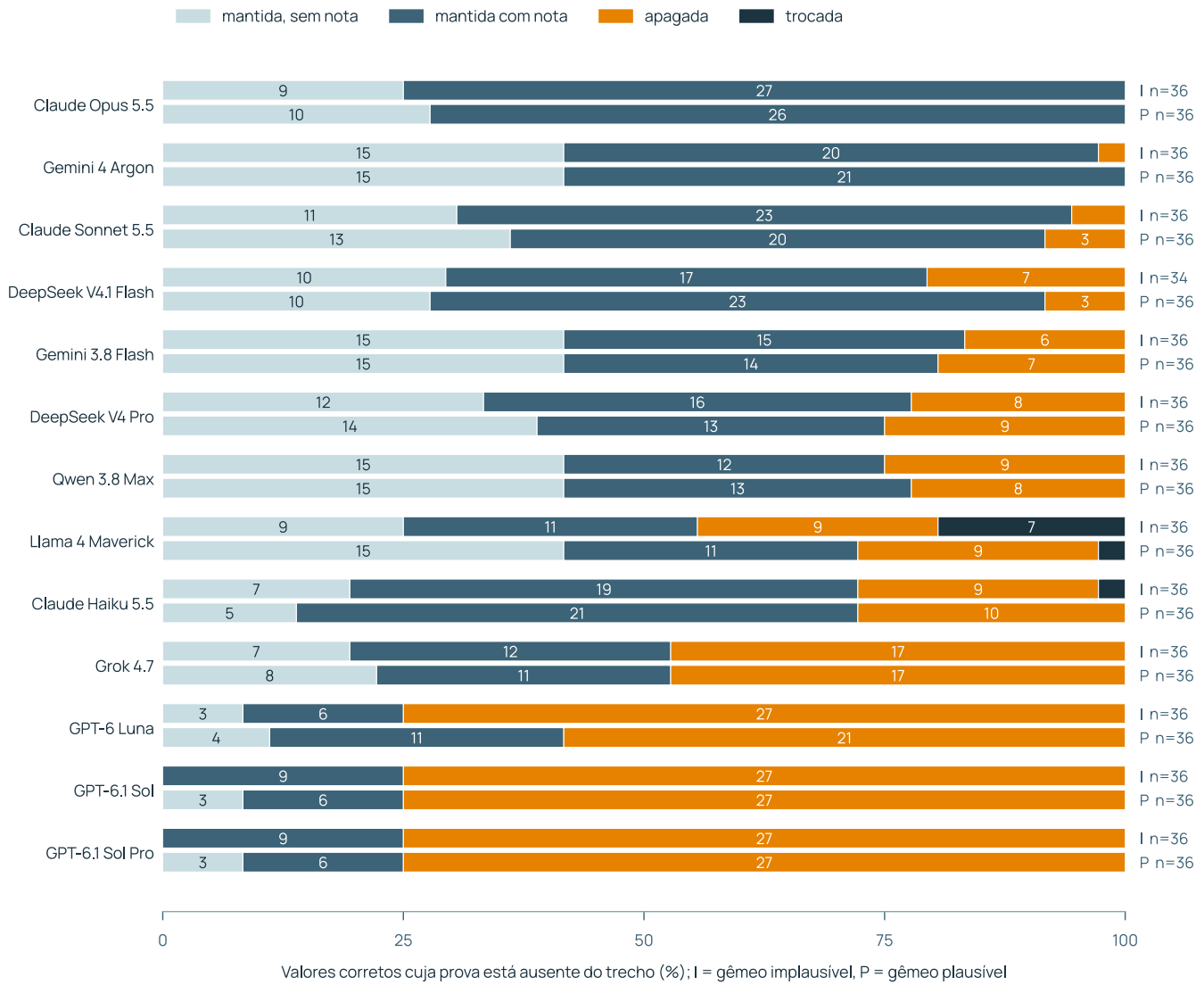
menos em todos os modelos.

3.4 Comportamento com o trecho: manter, sinalizar ou apagar

Com a prova removida, os modelos se distribuíram entre dois extremos (Figura 3). O Claude Opus 5.5 apagou 0 de 36 células em cada gêmeo e manteve com nota 27 e 26 delas. O Gemini 4 Argon apagou 1 de 36 no gêmeo implausível e 0 de 36 no plausível, com 20 e 21 células mantidas com nota. O Claude Sonnet 5.5 apagou 2 e 3. No outro extremo, o GPT-6.1 Sol e o GPT-6.1 Sol Pro apagaram 27 de 36 em cada gêmeo (75,0%; IC 95% 58,9% a 86,2%), o GPT-6 Luna, 27 e 21, e o Grok 4.7, 17 e 17. Os outros seis modelos ficaram entre 3 e 10 células apagadas por gêmeo.

As alterações colaterais acompanharam o apagamento (Figura 4). No laudo completo, todos os modelos ficaram em 0,2 por resposta ou menos. No trecho, o Claude Opus 5.5, o Gemini 4 Argon e o Claude Sonnet 5.5 alteraram 0,0, 0,1 e 0,2 células não alvo por resposta; o Grok 4.7, 3,2; o GPT-6 Luna, 3,8; o GPT-6.1 Sol e o Sol Pro, 5,0. Essas alterações foram, em sua maioria, a substituição de valores corretos de células não alvo por marcadores de ausência. Os erros plantados que continuavam verificáveis no trecho foram corrigidos por todos os modelos, com 1 omissão em 108 no Claude Sonnet 5.5 e no GPT-6 Luna, 2 em 108 no Claude Haiku 5.5 e nenhuma nos demais modelos (104 erros verificáveis no DeepSeek V4.1 Flash).

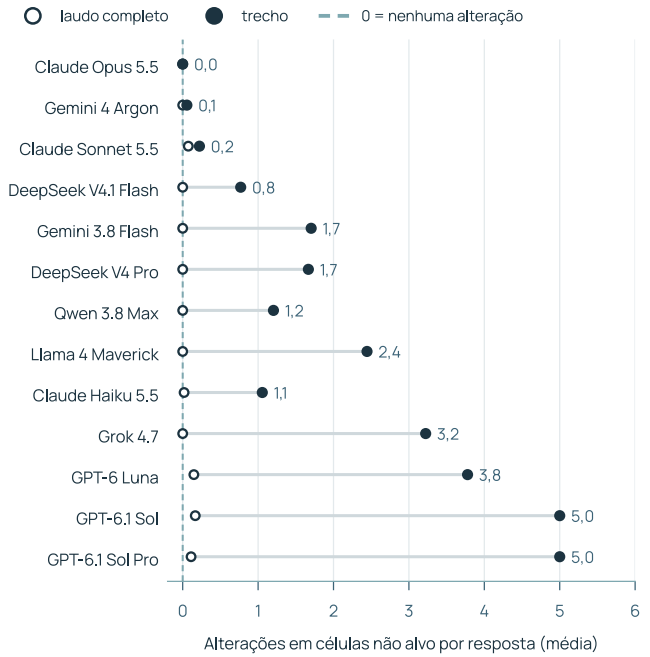
Figura 3. Comportamento do revisor sob evidência parcial: manter, sinalizar, apagar ou trocar



Desfecho de cada célula correta cuja prova está ausente do trecho, em barra de 100% por modelo e gêmeo (I, implausível; P, plausível), somadas as condições neutra, alerta e citação (n=36 células por barra; 34 para o DeepSeek V4.1 Flash no gêmeo implausível). Números nas barras são contagens de células (omitidos abaixo de 6%). Modelos ordenados pela proporção apagada. Azul-claro, mantida sem nota; azul-acinzentado, mantida com nota; âmbar, apagada; azul-escuro, trocada.

Fonte: Elaborado pelo autor (2026).

Figura 4. Alterações colaterais: mudanças em células que não estavam em revisão



Média de alterações em células não alvo, todas corretas como impressas, por resposta, no laudo completo (círculo vazio) e no trecho (círculo cheio), condições neutra, alerta e citação; 27 respostas por modelo e contexto (25 para o DeepSeek V4.1 Flash no trecho). A linha tracejada marca zero.

Fonte: Elaborado pelo autor (2026).

Figura 5. Efeito da instrução sobre o apagamento

	sem regra de preservação			com a regra
	neutra	alerta	alerta + citação	regra de preservação
Claude Opus 5.5	0/24	0/24	0/24	0/24
Gemini 4 Argon	1/24	0/24	0/24	0/24
Claude Sonnet 5.5	0/24	5/24	0/24	0/24
DeepSeek V4.1 Flash	2/23	6/23	2/24	0/24
Gemini 3.8 Flash	6/24	6/24	1/24	0/24
DeepSeek V4 Pro	5/24	6/24	6/24	0/24
Qwen 3.8 Max	6/24	6/24	5/24	0/24
Llama 4 Maverick	6/24	6/24	6/24	6/24
Claude Haiku 5.5	7/24	6/24	6/24	5/24
Grok 4.7	12/24	12/24	10/24	0/24
GPT-6 Luna	16/24	15/24	17/24	12/24
GPT-6.1 Sol	18/24	18/24	18/24	0/24
GPT-6.1 Sol Pro	18/24	18/24	18/24	0/24

Células corretas sem prova apagadas sobre o total de células sem prova, somados os dois gêmeos (n=24 por célula do mapa; 23 para o DeepSeek V4.1 Flash em duas condições), trecho com procedência declarada, por condição. Os cabeçalhos separam as três condições sem regra de preservação da condição com ela. A cor é proporcional à taxa. Modelos na ordem da Figura 3.

Fonte: Elaborado pelo autor (2026).

3.5 Teste de simetria

A taxa de apagamento da mesma célula foi igual ou quase igual nos dois gêmeos (Figura 3): em 11 dos 13 modelos, a diferença foi de no máximo uma célula. As duas exceções foram o GPT-6 Luna (27 contra 21) e o DeepSeek V4.1 Flash (7 de 34 contra 3 de 36). Nenhum modelo apagou mais no gêmeo implausível do que no plausível de forma consistente. O GPT-6.1 Sol apagou a frequência regional, a fração alélica paterna e a origem da variante, cujas frases de prova estavam fora do trecho, e manteve o sexo registrado, cujo valor continuava impresso no cabeçalho, embora sem explicação (9 de 9 respostas em cada gêmeo, no Sol e no Sol Pro; com nota em 9 de 9 no gêmeo implausível e em 6 de 9 no plausível).

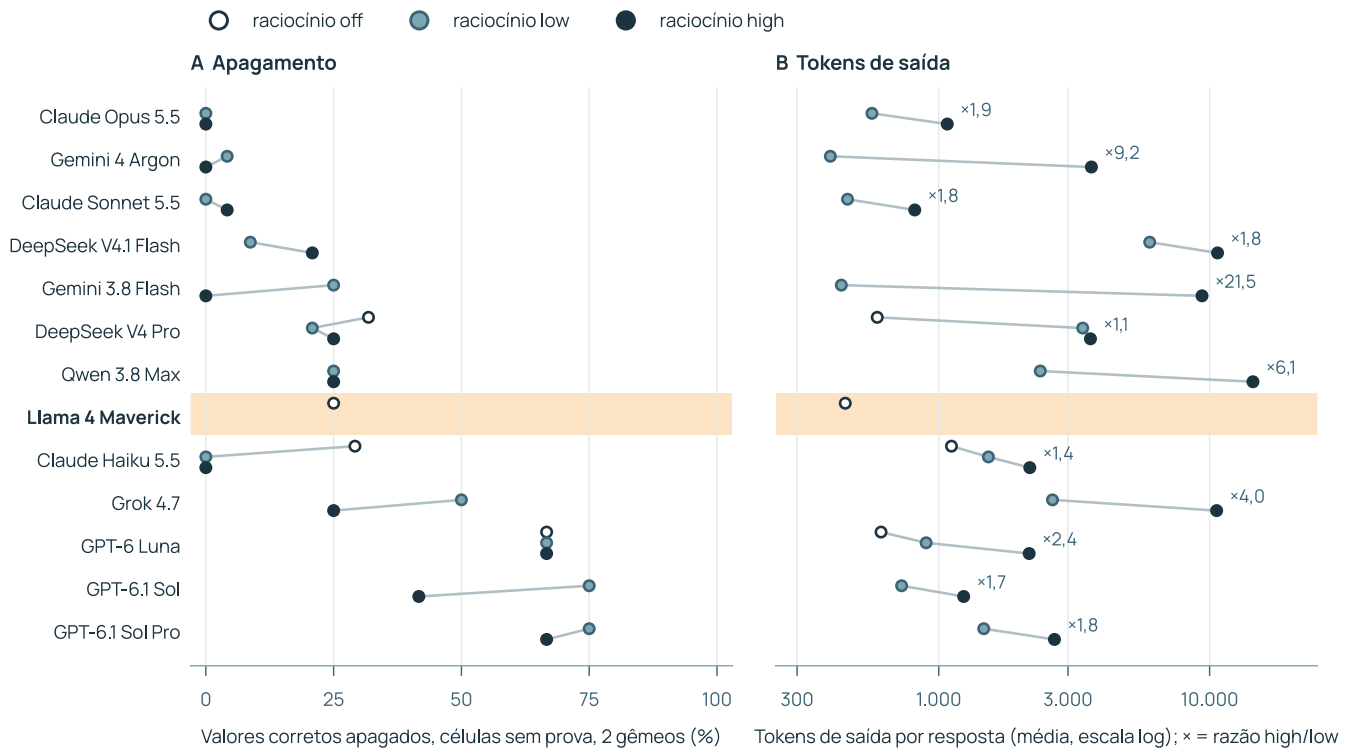
3.6 Efeito da instrução

As condições de alerta e de citação obrigatória não reduziram o apagamento em relação à condição neutra em 12 dos 13 modelos (Figura 5): o GPT-6.1 Sol apagou 18 de 24 células sem prova nas três condições, e o Grok 4.7, 12, 12 e 10. A exceção foi o Gemini 3.8 Flash, em que a citação obrigatória reduziu o apagamento de 6 de 24 para 1 de 24. Com a regra de preservação, o apagamento caiu a zero em 10 dos 13 modelos, incluindo GPT-6.1 Sol e Sol Pro (de 18 de 24 para 0 de 24), Grok 4.7, DeepSeek V4 Pro e Qwen 3.8 Max. Nesses modelos, as células que antes eram apagadas passaram a ser mantidas com nota, e as alterações colaterais do GPT-6.1 Sol e do Sol Pro caíram de 5,0 para 0,0 por resposta. O apagamento persistiu com a regra no GPT-6 Luna com raciocínio desligado (12 de 24), no Llama 4 Maverick (6 de 24) e no Claude Haiku 5.5 com raciocínio desligado (5 de 24).

3.7 Efeito do esforço de raciocínio

Na condição neutra do trecho, elevar o esforço de *low* para high zerou o apagamento no Gemini 3.8 Flash (de 6 de 24 para 0 de 24), reduziu-o pela metade no GPT-6.1 Sol (18 para 10) e no Grok 4.7 (12 para 6), e não o alterou de forma relevante no GPT-6.1 Sol Pro (18 para 16), no GPT-6 Luna (16 para 16), no Qwen 3.8 Max (6 para 6), no DeepSeek V4 Pro (5 para 6) e no DeepSeek V4.1 Flash (2 de 23 para 5 de 24) (Figura 6A). Ligar o raciocínio (off para *low*) zerou o apagamento no Claude Haiku 5.5 (7 para 0) e não o alterou no GPT-6 Luna (16 para 16) nem no DeepSeek V4 Pro (7 de 22 para 5 de 24). O esforço *high* multiplicou a média de *tokens* de saída por resposta por 1,1 (DeepSeek V4 Pro) a 21,5 (Gemini 3.8 Flash), com 1,7 no GPT-6.1 Sol e 4,0 no Grok 4.7 (Figura 6B).

Figura 6. Efeito do esforço de raciocínio sobre o apagamento e sobre o seu custo



Trecho com procedência declarada, condição neutra, 12 células sem prova por gêmeo e 18 respostas por nível. *Off*, raciocínio desligado onde a API permite (o Llama 4 Maverick, em âmbar claro, não tem o parâmetro); *low*, nível da rodada principal, exceto no Claude Haiku 5.5 e no GPT-6 Luna, em que a rodada principal foi *off*; high, rodada separada. (A) Proporção de células corretas sem prova apagadas, dois gêmeos somados (n=24). (B) Média de *tokens* de saída por resposta, escala logarítmica; o rótulo dá a razão high sobre *low*.

Fonte: Elaborado pelo autor (2026).

4 DISCUSSÃO

Modelos de linguagem com parâmetro de raciocínio, postos a revisar uma tabela extraída de um laudo genômico, não trocaram valores corretos que pareciam errados. Quando a prova do valor foi removida do material, alguns desses mesmos modelos apagaram o valor correto em três quartos das células, enquanto outros o mantiveram com uma nota. O risco medido, portanto, não é o de o revisor impor sua expectativa sobre o dado, e sim o de ele descartar um dado que não consegue conferir.

Estudos anteriores avaliaram LLMs como extratores de dados de notas clínicas, com acurácia alta em campos simples [5], e de fenótipos e de evidências para a classificação de variantes [27-29]; nesses estudos, o erro é um valor inventado ou omitido. A alucinação [19] e a concordância com premissas falsas [9,10] também foram descritas como um valor errado produzido pelo modelo, e não como a perda de um valor correto. Este estudo acrescenta três elementos: avalia o modelo como revisor de uma extração prévia, e não como extrator; separa, para o valor que não pode ser verificado, as condutas de manter, sinalizar e apagar; e controla a plausibilidade com gêmeos, o que distingue o apagamento por ausência de texto da troca por expectativa.

O resultado sobre o viés de plausibilidade difere do que a literatura de efeitos de conteúdo e de bajulação faria esperar. Modelos de linguagem erram mais em silogismos com conclusões implausíveis e concordam com premissas falsas apresentadas pelo usuário, inclusive em pedidos médicos [7-10]. Com a prova no documento e a instrução de verificar a tabela contra ele, 12 modelos trocaram 1 célula em 970, e o limite superior do intervalo, 0,58%, exclui taxas da ordem das relatadas em tarefas de concordância com o usuário. A troca sistemática do Llama 4 Maverick, que escreveu "*male*" numa paciente 46,XY mesmo com a explicação no laudo, é compatível com o que se descreve como efeito do conhecimento prévio do treinamento sobre a leitura do contexto [8,18]. A diferença entre esse modelo e os demais sugere que o parâmetro de raciocínio, ou o treinamento que o acompanha, protege contra a troca, mas o desenho não separa os dois.

O apagamento sob evidência parcial é um modo de falha que a literatura de abstenção e de fidelidade ao contexto não nomeia. Essa literatura trata do modelo que responde quando deveria abster-se, ou que segue o contexto quando deveria seguir o conhecimento prévio, e observa que as avaliações usuais recompensam responder [16,17,19,20]. O comportamento medido aqui é o inverso: o modelo tinha um valor correto, declarado como vindo do documento completo, e o removeu por não encontrá-lo no trecho. Essa conduta poderia ser lida como cautela, mas dois dados pesam contra. A declaração de procedência dizia ao modelo que a ausência no trecho era esperada, e a

sinalização com nota estava disponível como resposta. E o apagamento veio acompanhado de alterações em células não alvo corretas, 5,0 por resposta no GPT-6.1 Sol e no Sol Pro, que não têm leitura de cautela possível.

A simetria entre gêmeos indica o mecanismo. Se o apagamento dependesse da plausibilidade do valor, o gêmeo implausível seria sistematicamente mais apagado; em 11 dos 13 modelos, as taxas dos dois gêmeos diferiram em no máximo uma célula. O que decidiu o destino de uma célula foi a presença literal do valor no trecho: o sexo registrado, impresso no cabeçalho, não foi apagado por nenhum modelo, nem pelos que apagaram as outras três células em 27 de 27 oportunidades. Isso aproxima o comportamento de uma checagem de presença de texto e o afasta de um julgamento clínico [18,20].

A regra de preservação de uma frase zerou o apagamento em 10 dos 13 modelos e levou as alterações colaterais dos modelos GPT-6.1 a zero. O alerta verdadeiro e a citação obrigatória, ao contrário, não tiveram efeito. Isso é compatível com a literatura de seguimento de instruções, em que instruções explícitas são seguidas com mais regularidade do que exigências indiretas [30], e indica que o apagamento é um comportamento padrão desses modelos, não uma incapacidade de ler o material. Os três modelos que não obedeceram à regra, GPT-6 Luna e Claude Haiku 5.5 com raciocínio desligado e Llama 4 Maverick, são os que operaram sem raciocínio; no Claude Haiku 5.5, ligar o raciocínio bastou para a regra ser seguida.

Elevar o esforço de raciocínio foi menos eficaz e mais caro do que a instrução: o efeito variou por modelo e custou de 1,1 a 21,5 vezes mais *tokens* de saída, enquanto a instrução resolveu o problema em mais modelos e a custo fixo. Relatos de escala inversa com computação em tempo de teste e de explicações de raciocínio que não refletem a decisão do modelo mostram que mais raciocínio não implica mais fidelidade ao material [23,31].

Para laboratórios que usem uma ferramenta de extração com revisão automatizada, os resultados sugerem três recomendações, a serem tratadas como hipóteses a testar em laudos reais antes de orientar a prática. A instrução de revisão deve dizer o que fazer com um valor que não pode ser verificado, e a resposta esperada é manter e anotar. A avaliação de um revisor deve medir o apagamento e as alterações colaterais separadamente da acurácia das correções, porque modelos com 100% de correção dos erros plantados apagaram três quartos dos valores corretos sem prova. E a saída de um revisor automatizado não deve ser gravada no sistema sem conferência humana das remoções, dado que operadores tendem a aceitar sugestões de sistemas automatizados mesmo quando erradas [21,32,33].

No Brasil, esse cenário de uso é concreto: laboratórios de genética e de análises clínicas e os serviços de referência em doenças raras previstos na Política Nacional de Atenção Integral às Pessoas com Doenças Raras [34] emitem laudos que precisam ser transcritos para prontuários e sistemas de informação. A Política Nacional de Informação e Saúde Digital tem entre seus princípios a qualidade e a integridade dos dados em saúde e o uso seguro e responsável de tecnologias emergentes, e prevê, para a inteligência artificial, governança, supervisão humana e gestão de riscos [35]; o Programa SUS Digital tem por objetivo promover a transformação digital no Sistema Único de Saúde (SUS) e reconhece a Rede Nacional de Dados em Saúde (RNDS) como a plataforma digital de interoperabilidade do país [36], papel também definido em decreto [37]. Um revisor que apaga um valor correto antes que ele chegue ao prontuário ou à RNDS produz uma perda silenciosa, porque um campo vazio chama menos atenção do que um valor errado, o que contraria o princípio de qualidade e integridade dos dados. Esse risco é relevante num país em que a qualidade dos dados dos sistemas de informação em saúde não é acompanhada por um plano regular de avaliação [38] e em que a interoperabilidade entre esses sistemas ainda precisa ser promovida [39]. A conferência humana das remoções atende à exigência de supervisão humana da inteligência artificial [35]. Essas medidas, de baixo custo, devem ser testadas em laudos reais de laboratórios brasileiros antes de virar padrão.

O estudo tem limitações. Os resultados vêm de três laudos sintéticos, escritos para o estudo, de um único domínio; os valores contraintuitivos têm respaldo na literatura, mas cada laudo é um caso, e os achados precisam ser confirmados em laudos reais, de outros laboratórios e formatos, antes de orientar a prática. Há três repetições por célula, o que alarga os intervalos. O esforço *high* foi medido só na condição neutra. O trecho foi definido por uma única regra de corte; cortes mais agressivos não foram avaliados e podem alterar o comportamento dos modelos menores. A mensagem de sistema, herdada de uma versão do instrumento para revisões sistemáticas, descreve o documento como relato de ensaio clínico, e não como laudo genômico (Quadro 2); por ser idêntica em todas as chamadas, não afeta as comparações internas, mas pode ter alterado as taxas absolutas de troca e de apagamento, em direção que o desenho não permite estimar, e estudos futuros devem usar uma mensagem alinhada ao domínio genômico. O padrão “apagar é falha” é uma norma escolhida, defendida pela procedência declarada e pela simetria entre gêmeos. Por fim, os modelos avaliados mudam com frequência, e os resultados valem para as versões acessadas na data indicada.

5 CONCLUSÃO

Com a prova no material, modelos de linguagem com raciocínio preservam valores corretos que parecem errados. Sem a prova, o que distingue os modelos é se mantêm e sinalizam ou se apagam, e essa diferença é um padrão de comportamento que uma instrução explícita de preservação remove. Esses resultados vêm de laudos sintéticos de um único domínio; a recomendação de incluir a regra de preservação na instrução de revisão é uma hipótese a ser confirmada em laudos reais antes de orientar a prática. Estudos futuros devem testar o mesmo desenho em outros tipos de documento e com cortes de trecho de severidade variável, para medir em que ponto a instrução deixa de bastar.

Contribuição do autor. O autor foi responsável por todas as etapas do estudo: concepção e delineamento, desenvolvimento do instrumento e do corretor, coleta, análise e interpretação dos dados, redação e revisão do manuscrito e aprovação da versão final.

Agradecimentos. O autor agradece ao Laboratório Sabin, Brasília, DF, e, em especial, a Janete Ana Ribeiro Vaz e Sandra Santana Costa Soares, pelo incentivo constante à sua atividade de pesquisa.

Financiamento. O estudo foi custeado com recursos próprios do autor, sem financiamento externo ou institucional.

Disponibilidade de dados e materiais. Os prompts completos enviados aos modelos, os laudos e as tabelas sintéticos, o código do instrumento e do corretor, as respostas brutas e corrigidas e os dados de cada figura estão depositados em repositório público com DOI: https://github.com/gbbarra/llm-reviewer-discipline-eval; depósito no Zenodo, https://doi.org/10.5281/zenodo.23984451.

Conflitos de interesse. O autor declara vínculo empregatício com um laboratório clínico privado; este trabalho foi realizado de forma independente desse vínculo. O autor declara não possuir outros conflitos de interesse de natureza pessoal, profissional, institucional, comercial ou financeira relacionados a este manuscrito.

Declaração de uso de IA. Ferramentas de Inteligência Artificial foram utilizadas no desenvolvimento deste manuscrito para apoio à escrita do código do instrumento, do corretor e dos scripts que geram as figuras a partir dos dados, à redação e à revisão linguística do texto. As ferramentas utilizadas foram: Claude (Anthropic) e Gemini 3.8 Flash (Google). Esses assistentes incluem versões das famílias de modelos avaliadas. Os modelos avaliados são o objeto do estudo; as respostas foram classificadas por código determinístico, e cada número foi conferido contra os dados (seção 2.8). O autor declara que o uso dessas ferramentas não substituiu a autoria intelectual humana, não foi empregado como método científico autônomo e que assume total responsabilidade pelo conteúdo, integridade, originalidade e interpretações apresentadas neste trabalho.

REFERÊNCIAS

1. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol. 2017;17(1):152. doi: https://doi.org/10.1186/s12874-017-0431-4
2. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol. 2006;59(7):697-703. doi: https://doi.org/10.1016/j.jclinepi.2005.11.010
3. Gartlehner G, Kahwati L, Hilscher R, Thomas I, Kugley S, Crotty K, et al. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. Res Synth Methods. 2024;15(4):576-89. doi: https://doi.org/10.1002/jrsm.1710
4. Gartlehner G, Kugley S, Crotty K, Viswanathan M, Dobrescu A, Nussbaumer-Streit B, et al. Artificial intelligence-assisted data extraction with a large language model: a study within reviews. Ann Intern Med. 2025;178(12):1763-71. doi: https://doi.org/10.7326/annals-25-00739
5. Huang J, Yang DM, Rong R, Nezafati K, Treager C, Chi Z, et al. A critical assessment of using ChatGPT for extracting structured data from clinical notes. NPJ Digit Med. 2024;7(1):106. doi: https://doi.org/10.1038/s41746-024-01079-8
6. Zheng L, Chiang WL, Sheng Y, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: Advances in Neural Information Processing Systems 36. 2023. p. 46595-623. doi: https://doi.org/10.52202/075280-2020
7. Lampinen AK, Dasgupta I, Chan SCY, Sheahan HR, Creswell A, Kumaran D, et al. Language models, like humans, show content effects on reasoning tasks. PNAS Nexus. 2024;3(7):pgae233. doi: https://doi.org/10.1093/pnasnexus/pgae233
8. McCoy RT, Yao S, Friedman D, Hardy MD, Griffiths TL. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. Proc Natl Acad Sci U S A. 2024;121(41):e2322420121. doi: https://doi.org/10.1073/pnas.2322420121
9. Sharma M, Tong M, Korbak T, Duvenaud D, Askell A, Bowman SR, et al. Towards understanding sycophancy in language models. In: The Twelfth International Conference on Learning Representations (ICLR 2024); 2024 [citado 3 nov 2026]. Disponível em: https://arxiv.org/abs/2310.13548
10. Chen S, Gao M, Sasse K, Hartvigsen T, Anthony B, Fan L, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. NPJ Digit Med. 2025;8(1):605. doi: https://doi.org/10.1038/s41746-025-02008-z
11. Arcos-Burgos M, Muenke M. Genetics of population isolates. Clin Genet. 2002;61(4):233-47. doi: https://doi.org/10.1034/j.1399-0004.2002.610401.x
12. Hughes IA, Davies JD, Bunch TI, Pasterski V, Mastroyannopoulou K, MacDougall J. Androgen insensitivity syndrome. Lancet. 2012;380(9851):1419-28. doi: https://doi.org/10.1016/s0140-6736(12)60071-3

13. Campbell IM, Yuan B, Robberecht C, Pfundt R, Szafranski P, McEntagart ME, et al. Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. Am J Hum Genet. 2014;95(2):173-82. doi: https://doi.org/10.1016/j.ajhg.2014.07.003

14. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS recommendations for the description of sequence variants: 2016 update. Hum Mutat. 2016;37(6):564-9. doi: https://doi.org/10.1002/humu.22981

15. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015;17(5):405-24. doi: https://doi.org/10.1038/gim.2015.30

16. Rajpurkar P, Jia R, Liang P. Know what you don't know: unanswerable questions for SQuAD. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018. p. 784-9. doi: https://doi.org/10.18653/v1/p18-2124

17. Wen B, Yao J, Feng S, Xu C, Tsvetkov Y, Howe B, et al. Know your limits: a survey of abstention in large language models. Trans Assoc Comput Linguist. 2025;13:529-56. doi: https://doi.org/10.1162/tacl_a_00754

18. Xu R, Qi Z, Guo Z, Wang C, Wang H, Zhang Y, et al. Knowledge conflicts for LLMs: a survey. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024. p. 8541-65. doi: https://doi.org/10.18653/v1/2024.emnlp-main.486

19. Kalai AT, Nachum O, Vempala SS, Zhang E. Evaluating large language models for accuracy incentivizes hallucinations. Nature. 2026;653(8116):1047-51. doi: https://doi.org/10.1038/s41586-026-10549-w

20. Zhang Z, Peng X, Sun J, Yao R, Jiang B, Han B. Rethinking faithfulness in LLMs: a pairwise context-sensitive perspective [preprint]. arXiv; 2026 [citado 3 nov 2026]. Disponível em: https://arxiv.org/abs/2610.07894

21. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121-7. doi: https://doi.org/10.1136/amiajnl-2011-000089

22. Gardner M, Artzi Y, Basmov V, Berant J, Bogin B, Chen S, et al. Evaluating models' local decision boundaries via contrast sets. In: Findings of the Association for Computational Linguistics: EMNLP 2020. 2020. p. 1307-23. doi: https://doi.org/10.18653/v1/2020.findings-emnlp.117

23. Turpin M, Michael J, Perez E, Bowman S. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In: Advances in Neural Information Processing Systems 36. 2023. p. 74952-65. doi: https://doi.org/10.52202/075280-3275

24. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209-12. doi: https://doi.org/10.1080/01621459.1927.10502953

25. Hanley JA. If nothing goes wrong, is everything all right? Interpreting zero numerators. JAMA. 1983;249(13):1743. doi: https://doi.org/10.1001/jama.1983.03330370053031

26. Miller E. Adding error bars to evals: a statistical approach to language model evaluations [preprint]. arXiv; 2024 [citado 3 nov 2026]. Disponível em: https://arxiv.org/abs/2411.00640

27. Garcia BT, Westerfield L, Yelemali P, Gogate N, Rivera-Munoz EA, Du H, et al. Improving automated deep phenotyping through large language models using retrieval-augmented generation. Genome Med. 2025;17(1):91. doi: https://doi.org/10.1186/s13073-025-01521-w

28. Li S, Wang Y, Liu CM, Huang Y, Lam TW, Luo R. AutoPM3: enhancing variant interpretation via LLM-driven PM3 evidence extraction from scientific literature. Bioinformatics. 2025;41(7):btaf382. doi: https://doi.org/10.1093/bioinformatics/btaf382

29. Kim S, Zhou Y, Guo Y, Xiao C, Zheng K. Applying natural language processing and large language models to clinical notes for phenotyping and diagnosing rare diseases: a systematic review. J Am Med Inform Assoc. 2026;33(6):1225-35. doi: https://doi.org/10.1093/jamia/ocag045

30. Zhou J, Lu T, Mishra S, Brahma S, Basu S, Luan Y, et al. Instruction-following evaluation for large language models [preprint]. arXiv; 2023 [citado 3 nov 2026]. Disponível em: https://arxiv.org/abs/2311.07911

31. Gema AP, Hägele A, Chen R, Arditi A, Goldman-Wetzler J, Fraser-Taliente K, et al. Inverse scaling in test-time compute. Trans Mach Learn Res. 2025 [citado 3 nov 2026]. Disponível em: https://arxiv.org/abs/2507.14417

32. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. Hum Factors. 2010;52(3):381-410. doi: https://doi.org/10.1177/0018720810376055

33. Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, et al. Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology. 2023;307(4):e222176. doi: https://doi.org/10.1148/radiol.222176

34. Brasil. Ministério da Saúde. Portaria nº 199, de 30 de janeiro de 2014. Institui a Política Nacional de Atenção Integral às Pessoas com Doenças Raras [citado 19 nov 2026]. Disponível em: https://bvsms.saude.gov.br/bvs/saudelegis/gm/2014/prt0199_30_01_2014.html

35. Brasil. Ministério da Saúde. Portaria GM/MS nº 12.301, de 27 de outubro de 2026. Institui a Política Nacional de Informação e Saúde Digital, no âmbito do Sistema Único de Saúde. Diário Oficial da União. 28 out 2026;Seção 1 Extra A: 48 [citado 19 nov 2026]. Disponível em: https://www.in.gov.br/web/dou/-/portaria-gm/ms-n-12.301-de-27-de-outubro-de-2026-740877901

36. Brasil. Ministério da Saúde. Portaria GM/MS nº 3.232, de 1º de março de 2024. Institui o Programa SUS Digital. Diário Oficial da União. 4 mar 2024;Seção 1:52 [citado 19 nov 2026]. Disponível em: https://www.in.gov.br/web/dou/-/portaria-gm/ms-n-3.232-de-1-de-marco-de-2024-546278935

37. Brasil. Decreto nº 12.560, de 23 de julho de 2025. Dispõe sobre a Rede Nacional de Dados em Saúde e sobre as Plataformas SUS Digital. Diário Oficial da União. 24 jul 2025 [citado 19 nov 2026]. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2025/decreto/d12560.htm

38. Lima CRA, Schramm JMA, Coeli CM, Silva MEM. Revisão das dimensões de qualidade dos dados e métodos aplicados na avaliação dos sistemas de informação em saúde. Cad Saude Publica. 2009;25(10):2095-109. doi: https://doi.org/10.1590/s0102-311x2009001000002

39. Costa MVS, Camargos MCS, Viana SMN, Mendes UVS. Avanços e desafios da interoperabilidade no Sistema Único de Saúde. J Health Inform. 2025;17:1112. doi: https://doi.org/10.59681/2175-4411.v17.2025.1112

NOTAS EDITORIAIS

O Journal of Health Informatics é uma publicação científica mantida pela Sociedade Brasileira de Informática em Saúde, associação científica sem fins lucrativos dedicada ao desenvolvimento da Informática em Saúde e da Saúde Digital no Brasil. O processo editorial é conduzido com independência científica, sem interferência institucional, comercial ou de patrocinadores, e segue as boas práticas internacionais de ética em publicação científica.

Este artigo foi avaliado por pares, em sistema duplo-cego, por especialistas da área. A equipe editorial usou ferramentas de inteligência artificial, com supervisão humana, na triagem inicial, na conferência do texto, das referências e dos números e, quando necessário, no redesenho de figuras, quadros e tabelas, aprovado pelos autores na prova. As ferramentas e as salvaguardas estão na Política de uso de inteligência artificial da revista. As decisões editoriais foram tomadas por pessoas.

SBIS Sociedade Brasileira de Informática em Saúde