Journal of Health Informatics

J Health Inform. 2026;18:e1886
DOI: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1886

ARTIGO COMPLETO
CBIS 2026

Inteligência artificial generativa na governança de dados em saúde: estudo comparativo com recuperação documental

Generative artificial intelligence in health data governance: a comparative study with document retrieval

Gilberto Sussumu Hida[1] ✉, Erika Yahata[1], Carlos Alexandre Negri[1], Birajara Soares Machado[1]

ORCID: 0009-0003-2310-5323 · 0009-0007-8531-8446 · 0009-0001-9791-1227 · 0000-0001-7119-4170

[1] Einstein Hospital - E.Gate, São Paulo (SP), Brasil.

FICHA DO ARTIGO

TIPO
Artigo completo · CBIS 2026

IDENTIFICADOR
e1886

DOI
10.59681/2175-4411.v18.iEspecial.2026.1886

RECEBIDO
17/05/2026

APROVADO
20/07/2026

PROCESSO DE AVALIAÇÃO
Double Blind Review

EDITOR RESPONSÁVEL
Juliano de Souza Gaspar

ACESSO
Aberto, CC BY 4.0

AUTOR CORRESPONDENTE
Gilberto Sussumu Hida
gilberto.hida@einstein.br

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: Analisar o uso de Retrieval-Augmented Generation (RAG) para a geração de respostas sobre governança de dados em saúde. **Método:** Estudo comparativo com 30 perguntas e 60 respostas, geradas em duas condições: com uso e sem uso do RAG. As respostas foram avaliadas, em procedimento duplo-cego, por dois especialistas e um LLM judge, nas dimensões acurácia, completude, aderência, utilidade e clareza. Foram ajustados cumulative link mixed models para as dimensões ordinais e modelo logístico multinomial para a preferência global. **Resultados:** Os dois especialistas humanos atribuíram avaliações mais altas às respostas com RAG, especialmente em completude, aderência e utilidade. O LLM judge não reproduziu esse padrão e, em clareza, favoreceu a solução sem RAG. **Conclusão:** O RAG mostrou potencial para apoiar respostas mais úteis e aderentes à governança de dados em saúde, contribuindo para a compreensão de boas práticas e políticas voltadas ao equilíbrio entre utilidade, segurança e privacidade.

Descritores: governança em saúde; inteligência artificial generativa; modelos de linguagem de grande escala.

ABSTRACT

Objective: To analyze the use of Retrieval-Augmented Generation (RAG) for generating responses on health data governance. **Method:** A comparative study with 30 questions and 60 responses, generated under two conditions: with and without the use of RAG. The responses were evaluated, in a double-blind procedure, by two specialists and an LLM judge across the dimensions of accuracy, completeness, adherence, usefulness, and clarity. Cumulative link mixed models were fitted for the ordinal dimensions, and a multinomial logistic model was used for overall preference. **Results:** The two human specialists assigned higher ratings to the RAG-enabled responses, especially for completeness, adherence, and usefulness. The LLM judge did not replicate this pattern and, for clarity, favored the responses without RAG. **Conclusion:** RAG showed potential to support more useful and guideline-adherent responses in health data governance, contributing to the understanding of good practices and policies aimed at balancing utility, security, and privacy.

Keywords: health governance; generative artificial intelligence; large language models.

CBIS 2026 XXI CONGRESSO BRASILEIRO DE INFORMÁTICA EM SAÚDE

Artigo apresentado no XXI Congresso Brasileiro de Informática em Saúde (CBIS 2026), realizado de 23 a 25 de setembro de 2026, em Brasília (DF), e publicado na seção CBIS 2026 do Journal of Health Informatics.

COMO CITAR ESTE ARTIGO: Hida GS, Yahata E, Negri CA, Machado BS. Inteligência artificial generativa na governança de dados em saúde: estudo comparativo com recuperação documental. J Health Inform 2026;18:e1886. doi: https://doi.org/10.59681/2175-4411.v18.iEspecial.2026.1886

1 INTRODUÇÃO

O uso de tecnologias digitais na área da saúde, como prontuários eletrônicos, dispositivos médicos e aplicativos móveis, tem contribuído para o aumento no volume de dados gerados, caracterizando o fenômeno do *big data* em saúde [1]. A análise desses dados pode favorecer o avanço de pesquisas, possibilitando a melhoria de diagnósticos e tratamentos, além de fornecer subsídios para a formulação e o aprimoramento de políticas públicas [1].

Entretanto, o gerenciamento de grandes volumes de dados em saúde apresenta desafios como a garantia de segurança da informação e privacidade dos pacientes, a fim de evitar o uso indevido e a exposição de dados sensíveis [2]. Nesse contexto, a governança de dados se apresenta como um conjunto de práticas, políticas e diretrizes que contribuem para o equilíbrio entre utilidade, segurança e privacidade [1]. Entre essas práticas, destacam-se a definição de políticas e mecanismos de controle de acesso, a avaliação de aspectos legais e éticos para o uso responsável dos dados, a gestão do ciclo de vida dos dados, que abrange como os dados devem ser coletados, armazenados, compartilhados e descartados, assim como a implementação de mecanismos de auditoria e monitoramento [1].

Para a implementação de uma governança de dados adequada ao contexto de cada país, é necessário compreender as normas, leis e políticas nacionais que regulam o uso e o compartilhamento de dados, uma vez que não existe um modelo único de governança aplicável globalmente [3,4]. Dessa forma, mostra-se importante a capacitação e a clareza na aplicação de políticas e práticas de governança por parte dos envolvidos, incluindo profissionais de saúde e gestores, a fim de garantir o adequado funcionamento do ecossistema da Saúde Digital [4].

Porém, a grande quantidade de informações regulatórias provenientes de diversas fontes pode constituir um desafio para as partes envolvidas, sobretudo no que se refere à identificação e compreensão de requisitos aplicáveis em diferentes contextos, tais como regras de consentimento para coleta de dados, privacidade, segurança e conformidade no uso de dados [5]. Assim como na governança de dados, outras áreas, como genômica, pesquisa de novos medicamentos e apoio a diagnóstico, enfrentam desafios decorrentes do grande volume e da complexidade dos dados e informações disponíveis [6]. Nesse contexto, o uso de modelos de linguagem de grande escala (*Large Language Models*, LLMs) tem se apresentado como uma abordagem promissora para auxiliar na organização, interpretação e utilização desses dados [6-8].

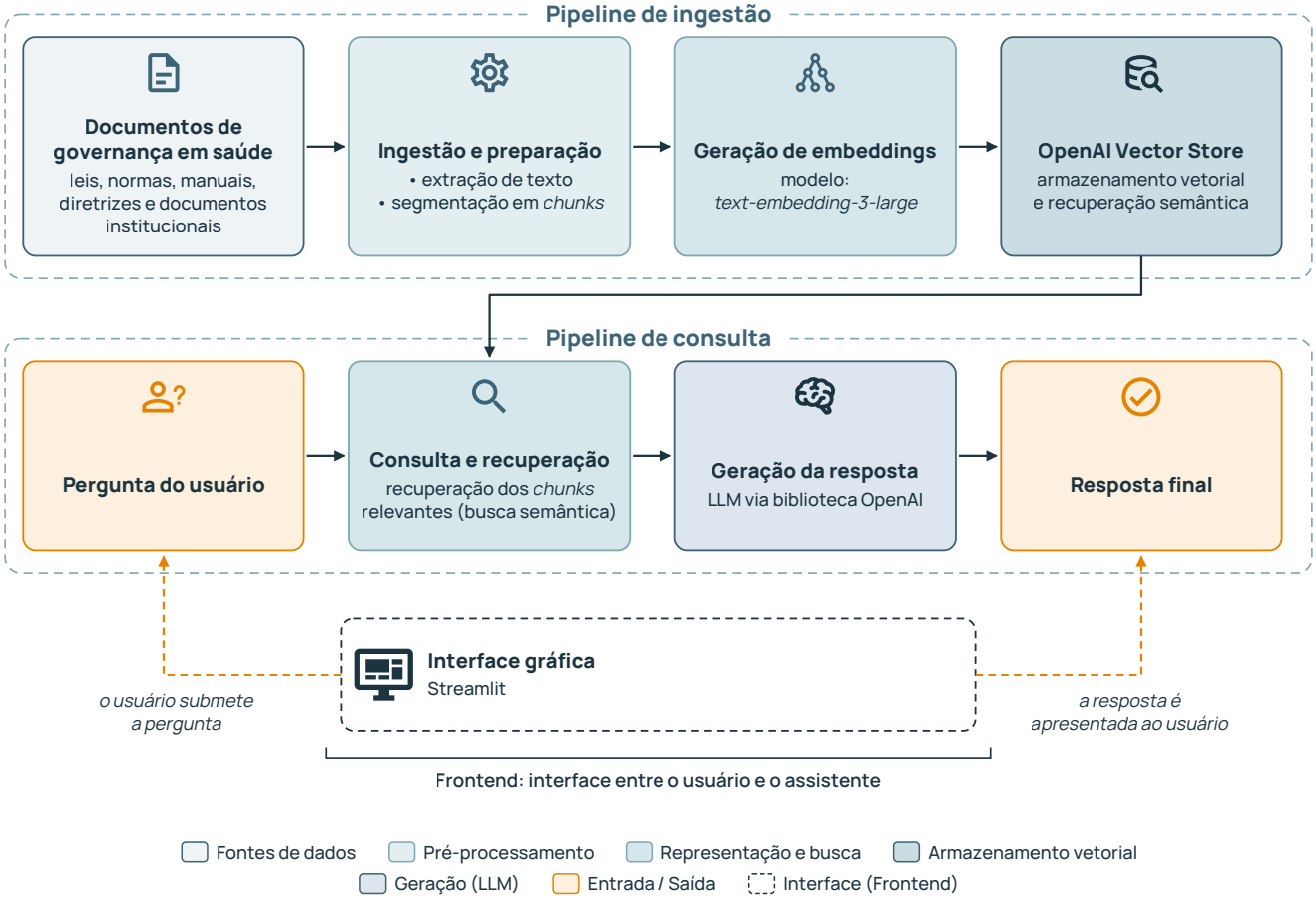
LLMs são modelos probabilísticos baseados em *deep learning*, capazes de gerar textos coerentes e apoiar tarefas de recuperação, síntese e interpretação de grandes volumes de informação textual [7]. Contudo, podem produzir conteúdos incorretos ou não fundamentados, fenômeno conhecido como "alucinação" [8]. Para mitigar essas limitações, têm sido investigadas estratégias como o *fine-tuning* e o *Retrieval-Augmented Generation* (RAG); porém, enquanto o *fine-tuning* pode demandar maior custo computacional, o RAG se apresenta como alternativa ao integrar recuperação semântica, *embeddings* e bases vetoriais para ampliar a confiabilidade das respostas [6-8].

Estratégias baseadas em RAG têm sido exploradas em diversos estudos, particularmente em contextos como procedimentos pré-operatórios, suporte à decisão clínica, oncologia e neurologia [9,10]. No entanto, ainda há evidências limitadas quanto à aplicação do RAG na governança de dados em saúde. Nesse contexto, o presente estudo adotou um desenho comparativo com o objetivo de avaliar se o uso de uma arquitetura RAG [11] agrega valor prático à geração de respostas no domínio de governança de dados em saúde. Além disso, a arquitetura RAG implica maior complexidade técnica e operacional quando comparada ao uso direto de modelos disponíveis comercialmente (*Software as a Service*) [12] e sua adoção justifica-se apenas quando resulta em ganhos na qualidade das respostas, considerando o contexto da aplicação.

2 MÉTODOS

Para o presente estudo de avaliação de desempenho do RAG no apoio à busca e à compreensão de temas de governança de dados em saúde, compararam-se duas configurações de geração de respostas: uma sem recuperação de informação e outra com uso de RAG. Para a condição com RAG, foi desenvolvida uma arquitetura específica (Figura 1), alimentada por documentos normativos e técnicos da área [13]. Esses documentos foram submetidos à extração de texto, segmentação em *chunks* (blocos) e vetorização por meio do modelo *text-embedding-3-large* [14], sendo então armazenados no OpenAI Vector Store [14] para recuperação semântica. Durante a consulta, a pergunta do usuário, recebida por uma interface gráfica em Streamlit [15], foi utilizada para recuperar os trechos mais relevantes, os quais eram encaminhados ao modelo de linguagem gpt-5.1-chat-latest, acessado via biblioteca da OpenAI [14], para compor a resposta final. Na condição sem RAG, a resposta foi gerada diretamente pelo mesmo modelo (gpt-5.1-chat-latest), sem recuperação de contexto documental.

Figura 1. Fluxograma da arquitetura implementada, destacando os pipelines de ingestão e consulta do sistema RAG utilizado no estudo.



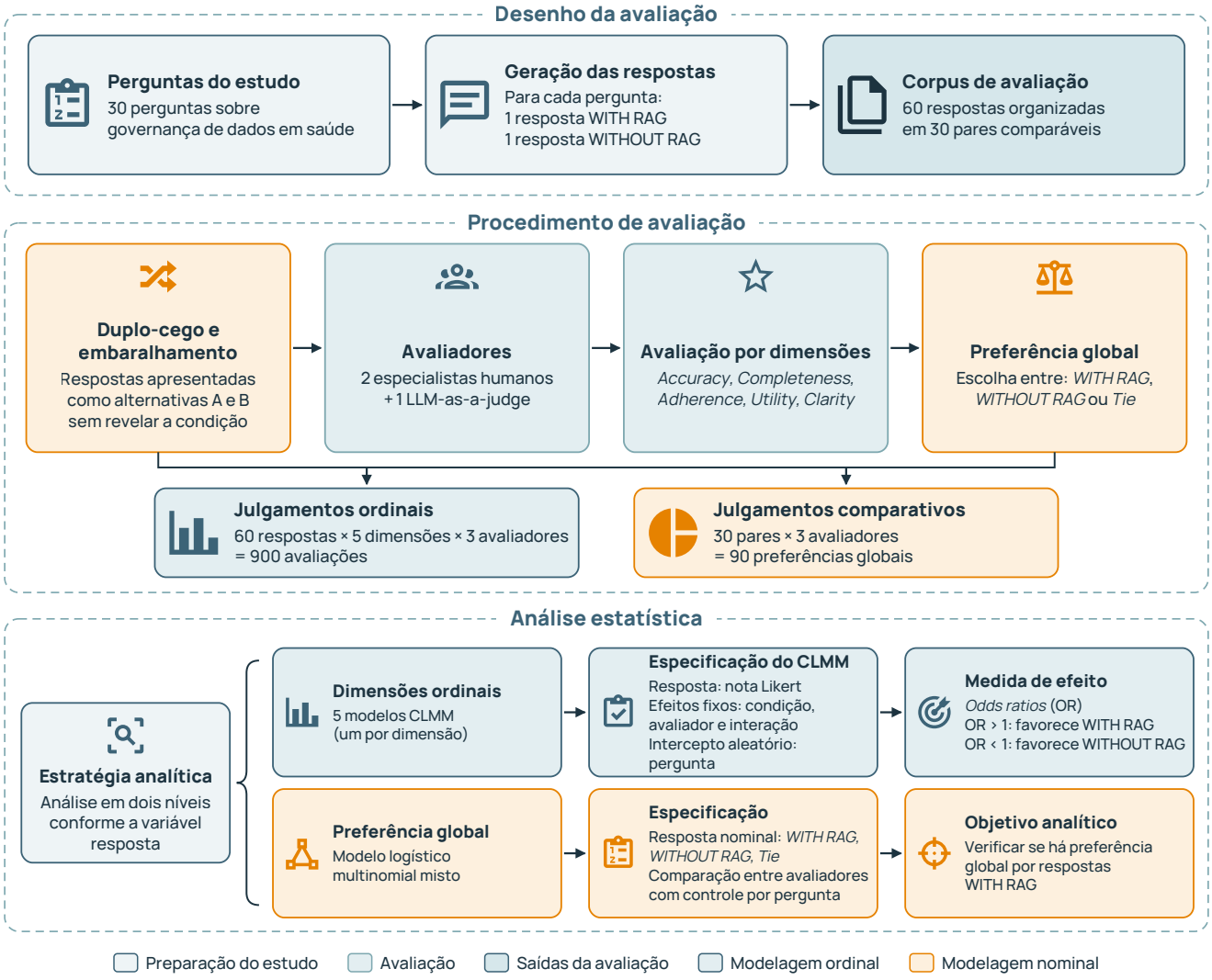
Fonte: Elaborado pelos autores (2026).

Elaborou-se um conjunto de 30 perguntas relacionadas à governança de dados no contexto da saúde [13], que foram aplicadas a ambas as arquiteturas, com e sem uso de RAG. As 60 respostas foram organizadas em 30 pares comparáveis, e o desempenho e a qualidade foram avaliados por dois avaliadores-autores (não participaram da elaboração das perguntas e respostas) com experiência em governança de dados em saúde e por um avaliador automatizado, no paradigma *LLM-as-a-judge* [16]. O avaliador automatizado utilizado foi o Claude Opus 4.6 (claude-opus-4-6), desenvolvido pela Anthropic [17] e acessado via API com temperatura igual a zero; os parâmetros top_p e seed não foram definidos, e cada julgamento correspondeu a uma única execução.

O processo seguiu um desenho duplo-cego, no qual cada avaliador recebeu uma planilha contendo as perguntas e as respostas geradas por ambas as arquiteturas, com e sem RAG. As respostas de cada par foram previamente randomizadas e apresentadas como alternativas A e B, impedindo a identificação da arquitetura correspondente. Esse procedimento visou reduzir o viés de expectativa e mitigar possíveis favorecimentos decorrentes do conhecimento prévio da condição experimental [18,19].

Cada resposta foi avaliada em cinco dimensões de qualidade: acurácia, completude, aderência, utilidade e clareza [20], aplicando a escala Likert (1 a 5). Além disso, para cada par de respostas referentes à mesma pergunta, os avaliadores registraram uma preferência entre o modelo com RAG (resposta gerada com recuperação de contexto), sem RAG (resposta gerada sem recuperação de contexto) ou empate (as respostas nas duas abordagens são equivalentes). Considerando as cinco dimensões de qualidade avaliadas por três avaliadores, o estudo totalizou 900 julgamentos ordinais, além de 90 julgamentos comparativos de preferência global entre os modelos. A Figura 2 apresenta o fluxograma da análise que foi conduzida.

Figura 2. Fluxograma do desenho de avaliação e da estratégia analítica adotados para comparar respostas geradas com e sem RAG.



Fonte: Elaborado pelos autores (2026).

A análise estatística foi estruturada em dois níveis, conforme a natureza da variável resposta. Para as cinco dimensões avaliadas em escala ordinal, ajustou-se um *cumulative link mixed model* (CLMM) [21], exclusivo para cada dimensão. Em cada modelo, a variável resposta foi a nota Likert atribuída e os efeitos fixos incluíram a condição experimental (com RAG versus sem RAG), o avaliador e a interação entre condição e avaliador. Cada CLMM incluiu 180 observações agrupadas em 30 perguntas. Os cinco ajustes produziram estimativas e erros-padrão finitos, sem avisos de não convergência; os diagnósticos adicionais são apresentados no material suplementar [13].

Para considerar a correlação entre observações provenientes da mesma pergunta, foi incluído um intercepto aleatório por pergunta, permitindo ao modelo CLMM capturar a dependência entre respostas associadas a uma mesma pergunta. Os coeficientes foram exponenciados e apresentados como *odds ratios* (OR), de modo que valores maiores que 1 indicam associação entre a condição com RAG e categorias mais altas da escala Likert, enquanto valores menores que 1 indicam associação com categorias mais baixas.

Para a variável de preferência global entre os modelos, composta por três categorias nominais (com RAG, sem RAG e empate), foi ajustado um modelo logístico multinomial [21], tendo o avaliador como preditor. Como análise de sensibilidade, ajustou-se adicionalmente um modelo multinomial com intercepto aleatório por pergunta, de modo a considerar o agrupamento das avaliações por item.

Os resultados foram apresentados com IC95% e nível de significância de 5%. A concordância foi estimada por kappas de Cohen ponderados quadraticamente nas dimensões ordinais, por kappas nominais na preferência global e pelos alfas de Krippendorff ordinal e nominal. A interpretação dos resultados priorizou não apenas a significância estatística, mas também a direção e magnitude dos efeitos estimados, considerando a natureza exploratória-aplicada da comparação entre respostas com e sem RAG em um domínio especializado.

Tabela 1. Desempenho por dimensão de avaliação e por avaliador em termos de odds ratio (OR). Como a categoria de referência do modelo foi a condição sem RAG, valores de OR > 1 indicam maior chance de a resposta com RAG receber avaliações mais altas, enquanto valores de OR < 1 indicam maior favorecimento à condição sem RAG.

Dimensão	OR (IC95%); p-valor		
	Avaliador 1	Avaliador 2	LLM *judge*
Acurácia	2,50 (1,00-6,24); 0,05	4,39 (1,61-11,93); <0,01	0,95 (0,37-2,45); 0,92
Completude	3,39 (1,33-8,63); 0,01	3,68 (1,41-9,64); <0,01	1,21 (0,43-3,37); 0,72
Aderência	1,93 (0,77-4,88); 0,16	7,39 (2,53-21,58); <0,01	1,49 (0,57-3,86); 0,41
Utilidade	2,09 (0,86-5,09); 0,11	5,45 (1,92-15,46); <0,01	0,63 (0,22-1,76); 0,38
Clareza	2,69 (1,08-6,61); 0,03	2,78 (1,03-7,53); 0,04	0,11 (0,04-0,31); <0,01

Fonte: Elaborado pelos autores (2026).

O estudo não foi submetido ao Comitê de Ética em Pesquisa por se tratar de avaliação tecnológica interna, conduzida exclusivamente pelos próprios autores, sem recrutamento de participantes externos, sem intervenção com usuários, pacientes ou profissionais, e sem coleta ou tratamento de dados pessoais, sensíveis ou identificáveis.

3 RESULTADOS

3.1 Efeito do RAG nas Dimensões Ordinais de Qualidade

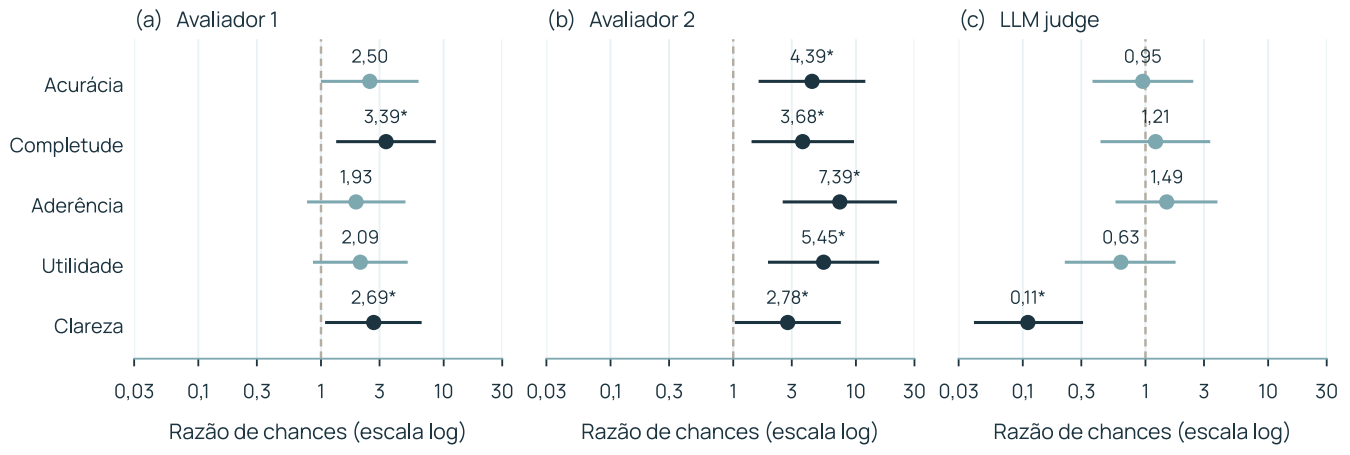
Em termos gerais, os resultados revelaram um padrão consistente de favorecimento ao modelo com uso do RAG entre os dois especialistas com experiência em governança de dados em saúde, Avaliador 1 e Avaliador 2; porém, esse padrão não foi reproduzido pelo LLM *judge* (Tabela 1).

Para o Avaliador 1, todas as estimativas apresentaram direção positiva, indicando associação entre o uso de RAG e avaliações mais elevadas em todas as dimensões analisadas (Figura 3.a). Esse efeito foi estatisticamente significativo para Completude (OR=3,39; IC95%: 1,33-8,63; p=0,01) e Clareza (OR=2,69; IC95%: 1,08-6,61; p=0,03). Em Acurácia, observou-se efeito positivo limítrofe (OR=2,50; IC95%: 1,00-6,24; p=0,05).

Para o Avaliador 2, o padrão foi ainda mais pronunciado (Figura 3.b). As cinco dimensões apresentaram associação positiva e estatisticamente significativa com a condição RAG. Os maiores efeitos foram observados em Aderência (OR=7,39; IC95%: 2,53-21,58; p<0,01) e Utilidade (OR=5,45; IC95%: 1,92-15,46; p<0,01), seguidos de Acurácia (OR=4,39; IC95%: 1,61-11,93; p<0,01), Completude (OR=3,68; IC95%: 1,41-9,64; p<0,01) e Clareza (OR=2,78; IC95%: 1,03-7,53; p=0,04). Em conjunto, esses resultados indicaram que, para o Avaliador 2, as respostas produzidas com RAG foram sistematicamente julgadas como superiores às respostas sem RAG em todas as dimensões analisadas.

Em contraste, o LLM *judge* exibiu comportamento substancialmente diferente (Figura 3.c). Em Acurácia (OR=0,95; IC95%: 0,37-2,45; p=0,92), Completude (OR=1,21; IC95%: 0,43-3,37; p=0,72) e Aderência (OR=1,49; IC95%: 0,57-3,86; p=0,41), os efeitos estimados foram menores e não alcançaram significância estatística. O resultado mais expressivo ocorreu em Clareza, dimensão em que o uso de RAG foi associado a avaliações significativamente mais baixas (OR=0,11; IC95%: 0,04-0,31; p<0,01). Enquanto as avaliações dos especialistas mostraram ganhos amplos com o uso de RAG, o juiz automatizado não apenas deixou de reproduzir esse padrão, como penalizou fortemente a clareza das respostas geradas com uso do RAG.

Figura 3. Resultados do CLMM; efeito do RAG nas avaliações ordinais, por dimensão e avaliador. (a) Avaliador 1. (b) Avaliador 2. (c) LLM judge.



Fonte: Elaborado pelos autores (2026).

Tabela 2. Resultados da análise com modelo logístico multinomial.

Contraste da resposta	Comparação	OR (IC95%)	p-valor	Interpretação
Com RAG vs Sem RAG	Avaliador 2 vs Avaliador 1	1,17 (0,34-4,06)	0,82	Sem diferença relevante entre avaliadores
Com RAG vs Sem RAG	LLM *judge* vs Avaliador 1	0,33 (0,11-1,03)	0,06	Tendência de menor preferência por RAG no LLM *judge*
Empate vs Sem RAG	Avaliador 2 vs Avaliador 1	5,00 (0,96-26,11)	0,06	Tendência de maior uso de empate pelo Avaliador 2
Empate vs Sem RAG	LLM *judge* vs Avaliador 1	1,00 (0,19-5,22)	1,00	Padrão semelhante ao Avaliador 1 nesse contraste

Fonte: Elaborado pelos autores (2026).

Em síntese, os modelos ordinais indicam que o efeito do RAG depende do perfil do avaliador. Sob julgamento dos especialistas, o RAG foi consistentemente associado a categorias mais altas da escala, especialmente nas dimensões de Completude, Aderência e Utilidade. Entretanto, sob julgamento automatizado, esse padrão não se sustentou, com estimativas próximas da ausência de associação na maior parte das dimensões e uma penalização expressiva em Clareza.

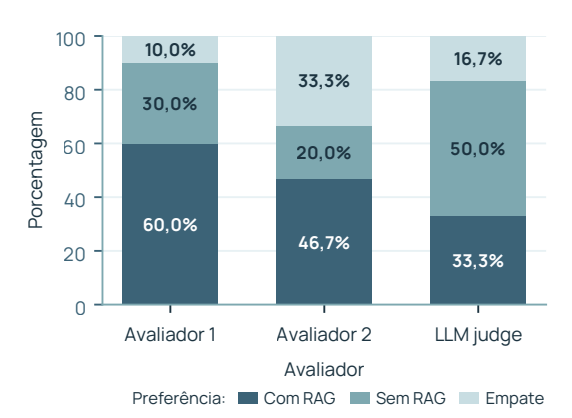
3.2 Concordância Entre Avaliadores

A concordância conjunta entre os três avaliadores foi baixa nas dimensões ordinais. Os coeficientes alfa de Krippendorff foram 0,186 (IC95%: −0,015 a 0,363) para Acurácia, 0,051 (−0,112 a 0,199) para Completude, 0,236 (0,042 a 0,413) para Aderência, 0,176 (−0,017 a 0,350) para Utilidade e 0,008 (−0,166 a 0,186) para Clareza. Os kappas ponderados pareados variaram de −0,001 a 0,456. Para a preferência global, o alfa de Krippendorff nominal foi 0,077 (−0,073 a 0,218), e os kappas nominais pareados variaram de 0,043 a 0,177. Os resultados completos são apresentados nas Tabelas Suplementares [13].

3.3 Preferência Global Entre Respostas

Para analisar a preferência global entre respostas com RAG, sem RAG e Empate, a análise descritiva demonstrou que as respostas produzidas com RAG foram as mais preferidas pelos especialistas, enquanto para o LLM *judge*, observou-se a predominância da preferência por respostas sem RAG, conforme ilustrado na Figura 4.

Figura 4. Resultados de preferência das respostas por avaliador (porcentagens empilhadas).



Fonte: Elaborado pelos autores (2026).

Para a análise inferencial, ajustou-se um modelo de regressão logística multinomial, adotando sem RAG como categoria de referência e o Avaliador 1 como avaliador de referência. O teste global do modelo, pela razão de verossimilhança (LR), indicou que a distribuição das preferências variou significativamente em função do avaliador (LR=10,69; df=4; p=0,03), sugerindo que a decisão final entre com RAG, sem RAG e Empate não foi homogênea entre os juízes. A Tabela 2 resume os achados.

Na comparação entre avaliadores, no contraste com RAG versus sem RAG, não se observou diferença relevante entre os dois avaliadores (Avaliador 2 vs Avaliador 1: OR=1,17; IC95%: 0,34-4,06; p=0,82), indicando um padrão semelhante entre ambos quanto à escolha direta entre respostas com e sem RAG. Em contraste, o LLM *judge* apresentou menor propensão a favorecer respostas com RAG em comparação ao avaliador humano de referência (LLM *judge* vs Avaliador 1: OR=0,33; IC95%: 0,11-1,03; p=0,06). Embora esse resultado não tenha atingido significância estatística no nível convencional de 5%, sua magnitude sugere tendência substantiva de menor favorecimento ao RAG por parte do avaliador automatizado.

No contraste Empate versus sem RAG, o Avaliador 2 apresentou maior propensão a registrar empate em comparação ao Avaliador 1 (OR=5,00; IC95%: 0,96-26,11; p=0,06), sugerindo um padrão de decisão mais cauteloso ou com maior frequência de situações em que não se observou superioridade suficientemente clara entre as respostas. Já o LLM *judge* não diferiu do Avaliador 1 nesse contraste (OR=1,00; IC95%: 0,19-5,22; p=1,00), indicando comportamento semelhante entre esses dois avaliadores quanto à relação entre empate e preferência por respostas sem RAG.

Para a análise de sensibilidade, ajustou-se também um modelo multinomial com intercepto aleatório por pergunta (question_id), de forma a considerar explicitamente o agrupamento das avaliações por item. Os coeficientes e erros-padrão obtidos foram praticamente idênticos aos do modelo principal. Mantiveram-se três achados centrais: a ausência de diferença relevante entre os dois avaliadores no contraste com RAG versus sem RAG; a menor propensão do LLM *judge* a favorecer respostas com RAG; e a maior tendência do Avaliador 2 a registrar empates. Essa estabilidade sugere que as conclusões

substantivas não foram alteradas pela introdução do agrupamento por pergunta.

Em síntese, a análise da preferência global indica que a escolha final entre respostas com RAG, sem RAG e Empate depende do perfil do avaliador. Os dois avaliadores apresentaram comportamentos relativamente próximos no contraste direto entre as respostas com e sem RAG, ao passo que o LLM *judge* mostrou-se menos inclinado a favorecer respostas com RAG. Além disso, o Avaliador 2 distinguiu-se por maior frequência relativa de empates.

4 DISCUSSÃO

A governança de dados tem papel fundamental no gerenciamento de dados do ecossistema da Saúde Digital, contribuindo para o equilíbrio entre utilidade, segurança e privacidade dos dados [1]. Porém, a busca e compreensão de informações regulatórias se apresentam como desafio [5] e, neste contexto, este estudo apresentou o uso de dois modelos baseados em Inteligência Artificial para reduzir o tempo de busca e auxiliar na compreensão das informações.

As respostas com RAG receberam avaliações mais altas dos especialistas, em concordância com estudo anterior no qual um modelo RAG apresentou maior acurácia que modelos sem RAG em 14 cenários clínicos (96,4% versus 92,9%) [9]. Além disso, o estudo comparou o tempo de resposta entre o modelo e profissionais da saúde para gerar instruções pré-operatórias, e os modelos levaram entre 15 e 20 segundos para as respostas, enquanto os profissionais levaram em média 10 minutos [9].

Tayebi Arasteh *et al.* [22] desenvolveram um modelo baseado em RAG para responder a questões na área de radiologia, avaliando cerca de 100 questões. Os resultados indicaram uma acurácia 54% superior aos modelos sem uso do RAG, sugerindo que abordagens baseadas em RAG podem apresentar melhor desempenho em contextos específicos [22]. Nossos resultados são consistentes com esses achados, ao indicar avaliações mais favoráveis às respostas com RAG no domínio de governança de dados em saúde. Em um domínio no qual decisões precisam ser justificáveis, alinhadas a normas e aplicáveis a rotinas institucionais, esse resultado é relevante, pois indica que a utilidade de uma arquitetura RAG deve ser apreciada à luz das necessidades do domínio, e não apenas de atributos textuais mais gerais.

Entre os especialistas, o uso de RAG esteve associado a avaliações mais altas de clareza e utilidade. De forma semelhante ao estudo [23], que avaliou o uso de RAG na interpretação de múltiplos guias regulatórios de alta complexidade na área farmacêutica e reportou resultados promissores no apoio à interpretação e à recuperação de informações normativas, nossos achados reforçam o potencial dessas abordagens para melhorar a qualidade, a relevância e a interpretabilidade das respostas em domínios regulatórios.

Do ponto de vista aplicado, os resultados trazem uma implicação importante para organizações de saúde e gestores de tecnologia: a avaliação de soluções assistidas por LLM, quando voltadas à governança de dados, não deve depender exclusivamente de julgamento automatizado. Neste estudo, observou-se divergência entre o julgamento dos avaliadores e o do LLM *judge*, inclusive na preferência global entre respostas com e sem RAG. Embora o caráter exploratório, o número reduzido de perguntas, a amplitude de alguns intervalos de confiança, o uso de um único LLM *judge* em uma configuração específica de modelo, *prompt* e corpus documental e o fato de os avaliadores humanos serem também coautores imponham cautela interpretativa, os resultados fornecem evidência inicial de que, em contextos regulatórios e especializados, a avaliação humana permanece necessária, especialmente nas etapas de validação, homologação e monitoramento de uso.

As análises de concordância reforçam essa cautela. De modo geral, a convergência entre os avaliadores foi limitada, sobretudo na preferência global, o que indica que a qualidade das respostas não foi percebida de maneira uniforme entre os juízes. Para a prática, isso significa que a implantação de um RAG em governança de dados na saúde deve ser acompanhada por um processo avaliativo estruturado, com critérios explícitos e participação de especialistas do domínio. Como agenda futura, seria importante ampliar o número de perguntas e avaliadores, comparar diferentes LLM *judges* e investigar modelos híbridos, nos quais o LLM possa apoiar a avaliação, mas não substituir, de forma isolada, o julgamento humano especializado.

5 CONCLUSÃO

Este estudo analisou o uso de uma arquitetura RAG no domínio da governança de dados em saúde, com o objetivo de verificar se a maior complexidade técnica e operacional dessa abordagem se justifica por ganhos na qualidade das respostas geradas. De modo geral, os resultados indicam que, sob avaliação humana especializada, o uso de RAG esteve associado ao melhor desempenho percebido, especialmente em dimensões relacionadas à completude, aderência e utilidade. Em contraste, esse padrão não foi reproduzido pelo LLM *judge*, que apresentou comportamento distinto tanto nas dimensões ordinais quanto na preferência global entre respostas com e sem RAG.

Em conjunto, os achados sugerem que o RAG é uma abordagem promissora para aplicações de governança de dados em saúde, particularmente quando a resposta precisa ser contextualizada, normativamente aderente e útil para especialistas do domínio. Ao mesmo tempo, os resultados também indicam que a avaliação da

qualidade dessas soluções não deve depender exclusivamente de julgamento automatizado, uma vez que o padrão observado entre os especialistas humanos não foi reproduzido de forma consistente pelo LLM *judge*. Assim, para fins de validação e adoção institucional, a participação de avaliadores humanos permanece um elemento relevante.

Como trabalhos futuros, recomenda-se ampliar o número de perguntas e de especialistas envolvidos, comparar diferentes LLM *judges*, investigar variações de *prompt* e explorar desenhos híbridos em que o LLM atue como apoio à avaliação, e não como substituta isolada do julgamento humano. Também será importante examinar se os resultados se mantêm em outras bases documentais, em diferentes configurações de RAG e em cenários reais de uso por equipes de governança, gestão da informação e Saúde Digital.

Contribuição dos autores. Os autores contribuíram igualmente em todas as etapas de construção do artigo: concepção e delineamento do estudo; coleta, análise e interpretação dos dados; redação e revisão crítica do manuscrito; e aprovação da versão final.

Agradecimentos. Agradecemos o apoio do Einstein Hospital - E.Gate.

Financiamento. Os autores não declararam fonte de financiamento.

Disponibilidade de dados e materiais. Os autores não declararam a disponibilidade dos dados e materiais da pesquisa.

Conflitos de interesse. Os autores não declararam conflitos de interesse.

Declaração de uso de IA. Foi utilizado o ChatGPT [14] como apoio à revisão textual do manuscrito e ao aprimoramento visual das figuras. Os autores revisaram integralmente o conteúdo final e assumem responsabilidade por sua precisão, originalidade e integridade científica.

REFERÊNCIAS

1. Abouelmehdi K, Beni-Hessane A, Khaloufi H. Big healthcare data: preserving security and privacy. J Big Data. 2018;5(1):1. doi: https://doi.org/10.1186/s40537-017-0110-7
2. Shaw J, Sekalala S. Health data justice: building new norms for health data governance. NPJ Digit Med. 2023;6(1):30. doi: https://doi.org/10.1038/s41746-023-00780-4
3. Bernier A, Molnár-Gábor F, Knoppers BM. The international data governance landscape. J Law Biosci. 2022;9(1):lsac005. doi: https://doi.org/10.1093/jlb/lsac005
4. Oktaviana R. S, Handayani PW, Hidayanto AN. Health organization challenges in health data governance implementation: A systematic review. JIPD. 2024;8(6):3892. doi: https://doi.org/10.24294/jipd.v8i6.3892
5. Winter JS, Davidson E. Harmonizing regulatory regimes for the governance of patient-generated health data. Telecommunications Policy. 2022;46(5):102285. doi: https://doi.org/10.1016/j.telpol.2021.102285
6. Zhou J, Li H, Chen S, Chen Z, Han Z, Gao X. Large language models in biomedicine and healthcare. npj Artif. Intell. 2025;1(1):44. doi: https://doi.org/10.1038/s44387-025-00047-1
7. Lin C, Kuo CF. Roles and potential of Large language models in healthcare: A comprehensive review. Biomed J. 2025;48(5):100868. doi: https://doi.org/10.1016/j.bj.2025.100868
8. Maity S, Saikia MJ. Large Language Models in Healthcare and Medical Applications: A Review. Bioengineering (Basel). 2025;12(6):631. doi: https://doi.org/10.3390/bioengineering12060631
9. Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia ATH, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. 2025;8(1):187. doi: https://doi.org/10.1038/s41746-025-01519-z
10. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. 2025;32(4):605-15. doi: https://doi.org/10.1093/jamia/ocaf008
11. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020;33:9459-74.
12. Xu X, Weytjens H, Zhang D, Lu Q, Weber I, Zhu L. Ragops: Operating and Managing Retrieval-Augmented Generation Pipelines. Elsevier BV [Preprint]. 2025. doi: https://doi.org/10.2139/ssrn.5319974
13. Hida GS. paper_RAG_data_governance [Internet]. GitHub; 2026 [citado 2026 abr 30]. Disponível em: https://github.com/sussumuedu/paper_RAG_data_governance.
14. OpenAI. API Platform [Internet]. 2026 [citado 2026 abr 28]. Disponível em: https://openai.com/api/.
15. Streamlit. Streamlit documentation [Internet]. [citado 2026 abr 28]. Disponível em: https://docs.streamlit.io/.
16. 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. San Diego, California, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS); 2023. p. 46595-623. doi: https://doi.org/10.52202/075280-2020
17. Anthropic. Introducing Claude Opus 4.6 [Internet]. 2026 [citado 2026 jul 28]. Disponível em: https://www.anthropic.com/news/claude-opus-4-6
18. Hróbjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. CMAJ. 2013;185(4):E201-11. doi: https://doi.org/10.1503/cmaj.120744

19. Shi L, Ma C, Liang W, Diao X, Ma W, Vosoughi S. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. In: Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: The Asian Federation of Natural Language Processing and The Association for Computational Linguistics; 2025. p. 292-314. doi: https://doi.org/10.18653/v1/2025.ijcnlp-long.18

20. Arora RK, Wei J, Hicks RS, Bowman P, Quiñonero-Candela J, Tsimpourlas F, et al. HealthBench: evaluating large language models towards improved human health. arXiv [Preprint]. 2025 [citado 2026 abr 28]. Disponível em: https://arxiv.org/abs/2505.08775.

21. Agresti A. Categorical Data Analysis. [s.l.]: Wiley; 2002. doi: https://doi.org/10.1002/0471249688

22. Tayebi Arasteh S, Lotfinia M, Bressem K, Siepmann R, Adams L, Ferber D, et al. RadioRAG: Online Retrieval-Augmented Generation for Radiology Question Answering. Radiol Artif Intell. 2025;7(4):e240476. doi: https://doi.org/10.1148/ryai.240476

23. Kim J, Hur M, Min M. From RAG to QA-RAG: Integrating Generative AI for Pharmaceutical Regulatory Compliance Process. In: Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing. New York, NY, USA: ACM; 2025. p. 1293-5. doi: https://doi.org/10.1145/3672608.3707749

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