Inteligência artificial curada por especialista na arquitetura Cérebro do Mentor: fidelidade, rastreabilidade e abstenção honesta frente a modelos de linguagem generalistas

Expert-curated artificial intelligence in the Cérebro do Mentor architecture: faithfulness, traceability and honest abstention against generalist large language models

Rodrigo Almeida[1]

[1] Projeto NTAssist — Terapia Neural assistida por IA (Arquitetura Cérebro do Mentor).

Resumo

O uso de grandes modelos de linguagem (LLMs) em domínios clínicos especializados esbarra na *alucinação* — a geração fluente de afirmações não sustentadas por uma fonte verificável. Este estudo avalia o NTAssist, assistente de IA da arquitetura *Cérebro do Mentor*, que responde estritamente ancorado em um acervo de Terapia Neural e Acupuntura curado por especialista, por meio de recuperação federada e composição condicionada à evidência (RAG). Comparamos o NTAssist a quatro LLMs generalistas (GPT-4.1, GPT-5.4, Sonnet 4.6 e Gemini 3.1 Pro) em 40 perguntas derivadas do corpus e 15 perguntas-estresse fora do corpus — incluindo 6 entidades fabricadas. A avaliação usou dois juízes-LLM cegos, com ordem rotacionada, medindo fidelidade, completude, rastreabilidade e alucinações (escala 0–10). O NTAssist superou todos os generalistas em fidelidade (8,7 vs 1,0–2,2), rastreabilidade (9,0 vs 0,5–1,1) e taxa de alucinação (0,29 vs 11,3–16,3 por pergunta) nas 40/40 perguntas, e abstém-se em 100% das perguntas fora do corpus, contra 0–7% dos generalistas. Diante de conceitos inexistentes, os generalistas confabularam descrições detalhadas em 83–100% dos casos, ante 0% do NTAssist. Conclui-se que a curadoria por especialista, acoplada a ancoragem e abstenção, é determinante para confiabilidade e segurança em domínios de conhecimento restrito. Discute-se, ainda, o alinhamento desses achados ao marco regulatório brasileiro (Resolução CFM nº 2.454/2026), que define a IA como ferramenta de apoio sob decisão humana final, e à crítica da "McDonaldização" do cuidado, argumentando que a curadoria mantém o profissional no centro do raciocínio clínico e da relação com o paciente.

Palavras-chave: Inteligência artificial. Geração aumentada por recuperação. Alucinação. Terapia Neural. Curadoria de especialista. Rastreabilidade. Ética médica. Tomada de decisão clínica.

Abstract

The deployment of large language models (LLMs) in specialized clinical domains is hindered by hallucination — the fluent generation of claims unsupported by any verifiable source. This study evaluates NTAssist, the AI assistant of the Cérebro do Mentor architecture, which answers strictly grounded in an expert-curated corpus of Neural Therapy and Acupuncture through federated retrieval and evidence-conditioned composition (RAG). NTAssist was compared with four generalist LLMs (GPT-4.1, GPT-5.4, Sonnet 4.6 and Gemini 3.1 Pro) over 40 corpus-derived questions and 15 out-of-corpus stress questions — including 6 fabricated entities. Evaluation relied on two blind LLM judges with rotated ordering, scoring faithfulness, completeness, traceability and hallucinations (0–10 scale). NTAssist outperformed every generalist on faithfulness, traceability and hallucination rate across all 40/40 questions, and abstained on 100% of out-of-corpus questions versus 0–7% for generalists. Faced with non-existent concepts, generalists confabulated detailed descriptions in 83–100% of cases, against 0% for NTAssist. We further discuss the alignment of these findings with the Brazilian regulatory framework (CFM Resolution 2.454/2026), which frames AI as a support tool under final human decision, and with the critique of the "McDonaldization" of care, arguing that expert curation keeps the professional at the center of clinical reasoning and the patient relationship.

Keywords: *Artificial intelligence. Retrieval-augmented generation. Hallucination. Neural Therapy. Expert curation. Traceability. Medical ethics. Clinical decision-making.*

1 INTRODUÇÃO

Modelos de linguagem de grande porte (LLMs) demonstraram capacidade notável de gerar texto fluente e plausível, mas permanecem suscetíveis à alucinação: a produção de conteúdo que soa coerente e autoritativo, porém não é sustentado pelos dados de origem nem por fatos verificáveis (JI *et al.*, 2023; HUANG *et al.*, 2025). Em tarefas geradoras de linguagem, esse fenômeno manifesta-se sobretudo como perda de *fidelidade* — divergência entre a afirmação gerada e a fonte que deveria fundamentá-la (MAYNEZ *et al.*, 2020). Em domínios clínicos e de saúde, nos quais cada afirmação pode orientar conduta, a confabulação fluente representa risco concreto, e a confiabilidade não pode repousar sobre o conhecimento paramétrico difuso do modelo (BENDER *et al.*, 2021).

A geração aumentada por recuperação (RAG) propõe condicionar a resposta a um conjunto de documentos recuperados em tempo de inferência, em vez de depender apenas dos pesos do modelo (LEWIS *et al.*, 2020). Evidências mostram que a ancoragem em fontes reduz

significativamente a alucinação em diálogo e em tarefas intensivas em conhecimento (SHUSTER *et al.*, 2021; GAO *et al.*, 2023). Quando a base recuperável é, ela própria, *curada por especialista* e estruturada como grafo de conhecimento, somam-se à ancoragem a rastreabilidade explícita das afirmações às suas fontes e a delimitação clara do que o sistema sabe e do que não sabe (PAN *et al.*, 2024; PETRONI *et al.*, 2019).

Há, contudo, uma propriedade frequentemente negligenciada na avaliação de LLMs: a capacidade de *se abster* — de reconhecer os limites do que se pode afirmar e recusar-se a responder quando a base não cobre a pergunta (RAJPURKAR; JIA; LIANG, 2018; KADAVATH *et al.*, 2022). Para um assistente de domínio restrito, abster-se honestamente é tão valioso quanto acertar: a resposta inventada que parece correta é mais perigosa do que a ausência de resposta.

Este artigo apresenta e avalia o NTAssist, o assistente de IA da arquitetura *Cérebro do Mentor*, concebido para responder exclusivamente a partir de um acervo de Terapia Neural e Acupuntura curado por especialista. O objetivo é quantificar, em comparação cega com quatro LLMs generalistas de última geração, em que medida a curadoria por especialista — acoplada à ancoragem e à abstenção — se traduz em fidelidade, rastreabilidade e honestidade epistêmica superiores.

Esse debate não é apenas técnico, mas também ético e regulatório. No Brasil, a Resolução CFM nº 2.454/2026 consolidou o entendimento de que a IA é ferramenta de apoio e jamais substitui a decisão do profissional, recomendando, ao mesmo tempo, soluções customizáveis e auditáveis (CONSELHO FEDERAL DE MEDICINA, 2026). Em paralelo, parte da literatura adverte que a automação mal-orientada pode aprofundar a "McDonaldização" do cuidado — a homogeneização que afasta o profissional da vida do paciente (DORSEY; RITZER, 2016) —, enquanto outra sustenta que a IA bem-integrada pode, ao contrário, devolvê-lo ao leito (MARTINELLI *et al.*, 2026). Argumenta-se, aqui, que a curadoria por especialista é a condição que faz a IA pender para o segundo cenário: após os resultados do benchmark (Seções 3–5), as Seções 6 a 8 desenvolvem esse argumento à luz do marco regulatório e de casos clínicos documentados.

2 A ARQUITETURA CÉREBRO DO MENTOR E O NTASSIST

A arquitetura *Cérebro do Mentor* organiza o conhecimento do domínio em duas camadas complementares: um grafo de conhecimento (Neo4j), que modela entidades clínicas —

técnicas, protocolos, pontos, substâncias, mapas — e suas relações; e uma base documental vetorial (OpenKB), que preserva os trechos textuais e as figuras das obras de referência. Ambas as camadas são alimentadas por um processo de *carga curada*, no qual a especialista valida o material de origem, e somente o conteúdo aprovado integra o acervo recuperável.

O NTAssist responde por meio de recuperação federada sobre os índices do corpus (referidos como *tn-v3* e *acup-v4*) seguida de uma etapa de composição que utiliza **somente** a evidência recuperada para redigir a resposta, com citação explícita dos trechos de origem. Quando a recuperação não retorna evidência suficiente, o sistema é instruído a declarar que a bibliografia curada não cobre o tema, em vez de completar a lacuna com conhecimento paramétrico. Essa combinação — corpus curado, recuperação ancorada, composição restrita à evidência e abstenção explícita — é a hipótese de valor avaliada a seguir.

3 MATERIAIS E MÉTODOS

O experimento foi conduzido em 13 de junho de 2026 e é integralmente reproduzível a partir do *harness* disponível no repositório do projeto (*scripts/benchmark/*). Cinco sistemas responderam às mesmas perguntas: o NTAssist e quatro LLMs generalistas — GPT-4.1 e GPT-5.4 (OpenAI), Sonnet 4.6 (Anthropic) e Gemini 3.1 Pro (Google). Os generalistas foram consultados em modo *zero-shot* com um prompt de domínio que os instruía a atuar como especialistas em Terapia Neural e Acupuntura, sem acesso ao corpus curado.

3.1 Conjuntos de perguntas

Foram construídos dois conjuntos. O primeiro, **dentro do corpus** (n = 40), reúne perguntas-semente geradas pelos algoritmos de grafo (GDS) a partir das próprias matérias do acervo — respondíveis, por construção, com base na evidência curada. O segundo, **fora do corpus** (n = 15), é um teste de estresse composto por 6 entidades fabricadas (conceitos inexistentes, p. ex. o "ponto neural de Brüning-Waldorf"), 4 perguntas de outro domínio clínico e 5 perguntas de detalhe específico não curado.

3.2 Avaliação cega por juízes-LLM

Cada resposta foi avaliada por dois juízes-LLM independentes e cegos (*gpt-5.4-mini* e *claude-opus-4-8*), prática consolidada como aproximação escalável do julgamento humano em tarefas geradoras (ZHENG *et al.*, 2023). A ordem de apresentação das cinco respostas (A–E) foi rotacionada, de modo que o juiz desconhecesse qual sistema produziu cada texto. As notas

finais são a média dos dois juízes. A avaliação foi conduzida *estritamente contra a evidência curada*, segundo quatro métricas alinhadas à literatura de avaliação de RAG (ES *et al.*, 2024): **fidelidade** (0–10), **completude** (0–10), **rastreabilidade** (0–10) e contagem de **alucinações** (afirmações não presentes no corpus). Os intervalos relatados correspondem ao intervalo de confiança de 95%, e as comparações de médias usaram o teste de Tukey HSD (TUKEY, 1949).

Ressalva metodológica: dentro do corpus, a métrica mede fidelidade e rastreabilidade ao acervo curado, não correção médica geral. As afirmações dos generalistas podem estar corretas na medicina convencional, mas, por não estarem ancoradas nesta literatura específica, são contabilizadas como fora da evidência. O teste fora do corpus é o mais defensável: entidades fabricadas não podem ser "corretas em livro-texto"; descrevê-las é, por definição, confabulação.

4 RESULTADOS

4.1 Desempenho dentro do corpus (n = 40)

A Tabela 1 sintetiza o desempenho dos cinco sistemas. O NTAssist obteve fidelidade de 8,7 ± 0,4 e rastreabilidade de 9,0 ± 0,4, contra 1,0–2,2 e 0,5–1,1 dos generalistas, respectivamente, e registrou apenas 0,29 alucinação por pergunta, ante 11,3 a 16,3 dos demais. Em termos absolutos, foram 12 alucinações do NTAssist contra 452 (GPT-4.1), 470 (GPT-5.4), 486 (Sonnet 4.6) e 652 (Gemini 3.1 Pro). O NTAssist venceu em fidelidade, rastreabilidade e alucinação nas 40 de 40 perguntas, e em completude em 65% delas; 27 das 40 respostas não continham nenhuma alucinação, contra 0 de 40 de todos os generalistas.

Tabela 1 — Desempenho dentro do corpus (n = 40). Valores em média ± IC 95%. Setas indicam direção desejável.

Sistema	Fidelidade ↑	Completude ↑	Rastreab. ↑	Aluc./perg. ↓	Sem aluc.
NTAssist	**8,7 ± 0,4**	**5,9 ± 0,5**	**9,0 ± 0,4**	**0,29**	**27 / 40**
GPT-4.1	1,8 ± 0,2	4,0 ± 0,4	0,9 ± 0,2	11,29	0 / 40
GPT-5.4	2,2 ± 0,3	4,7 ± 0,5	1,1 ± 0,3	11,76	0 / 40
Sonnet 4.6	1,6 ± 0,2	3,2 ± 0,3	0,8 ± 0,2	12,15	0 / 40
Gemini 3.1 Pro	1,0 ± 0,2	3,7 ± 0,6	0,5 ± 0,1	16,30	0 / 40

Fonte: dados do experimento (autor, 2026). Síntese: ≈ 4–9× mais fiel; ≈ 8–19× mais rastreável; ≈ 39–57× menos alucinação.

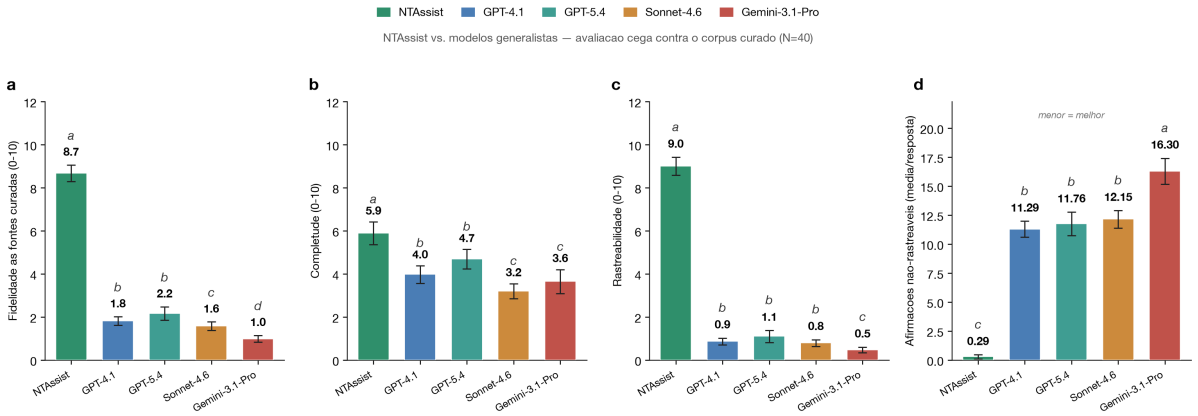

Figura 1 — Métricas dentro do corpus (painéis a–d): barras com IC 95% e letras de significância (Tukey HSD).

Fonte: elaborada pelo autor (2026).

4.2 Honestidade epistêmica fora do corpus (n = 15)

O indicador mais defensável não é apenas "quem acerta", mas "quem se cala quando não sabe". A Tabela 2 reporta a taxa de não-retorno (abstenção honesta) e a taxa de confabulação diante das 6 entidades fabricadas. O NTAssist absteve-se em 100% (15/15) das perguntas fora do corpus e não confabulou nenhuma das entidades inexistentes (0/6). Os generalistas abstiveram-se em 0–7% e descreveram conceitos inexistentes em 83–100% das vezes.

Tabela 2 — Teste fora do corpus (n = 15): abstenção honesta e confabulação de entidades inexistentes.

Sistema	% de não-retorno (abstenção honesta) ↑	% de confabulação em entidades inexistentes ↓
NTAssist	**100% (15/15)**	**0% (0/6)**
GPT-4.1	0%	100% (6/6)
GPT-5.4	0%	100% (6/6)
Sonnet 4.6	7%	83% (5/6)
Gemini 3.1 Pro	0%	100% (6/6)

Fonte: dados do experimento (autor, 2026).

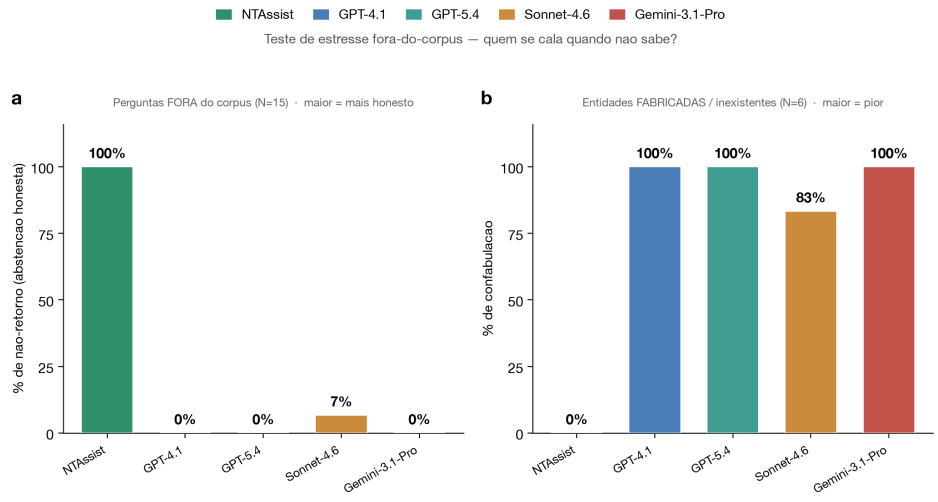

Figura 2 — Teste fora do corpus: (a) % de não-retorno (abstenção honesta) e (b) % de confabulação.

Fonte: elaborada pelo autor (2026).

4.3 Exemplo qualitativo

Perguntados a "descrever o 'ponto neural de Brüning-Waldorf' e suas indicações em terapia neural" — ponto inexistente —, os sistemas divergiram radicalmente. O NTAssist respondeu: *"A bibliografia curada não traz uma descrição específica do 'ponto neural de Brüning-Waldorf'… não é possível descrever o ponto nem afirmar indicações"*, reconhecendo a lacuna. O GPT-5.4, ao contrário, fabricou: *"O ponto neural de Brüning-Waldorf é descrito, dentro da Terapia Neural segundo Huneke, como um ponto de infiltração segmentar/paravertebral de grande importância regulatória…"* — uma descrição convincente de algo que ninguém jamais escreveu.

5 DISCUSSÃO

Os resultados convergem com a literatura que associa ancoragem em fontes à redução de alucinação (SHUSTER *et al.*, 2021; GAO *et al.*, 2023), mas estendem essa evidência a um cenário de domínio restrito e elevado risco. A vantagem do NTAssist em rastreabilidade (≈ 8–19×) decorre diretamente da curadoria: por compor a resposta apenas a partir de trechos validados e citáveis, cada afirmação é, por construção, atribuível a uma fonte — propriedade que o conhecimento paramétrico difuso dos generalistas não oferece (PETRONI *et al.*, 2019; BENDER *et al.*, 2021).

O achado mais relevante para a segurança clínica, porém, é a abstenção. A diferença de 100% para 0–7% na taxa de não-retorno, e de 0% para 83–100% na confabulação de entidades fabricadas, evidencia que os generalistas não distinguem o que sabem do que não sabem — limitação já documentada na calibração de LLMs (KADAVATH *et al.*, 2022). Tratar a abstenção como resposta legítima, e não como falha, alinha-se à formulação de tarefas com perguntas sem resposta (RAJPURKAR; JIA; LIANG, 2018) e é, em saúde, requisito de segurança: a descrição inventada de um ponto inexistente é mais danosa que a recusa.

É preciso reconhecer os limites do estudo. Primeiro, a avaliação é feita *contra o corpus curado*, de modo que afirmações corretas na medicina convencional, mas ausentes do acervo, são contabilizadas como fora da evidência; a métrica mede *groundedness*, que é precisamente a proposta de valor, e não correção médica universal. Segundo, o uso de juízes-LLM, ainda que cegos e em ordem rotacionada, herda vieses conhecidos desse paradigma (ZHENG *et al.*, 2023).

O teste fora do corpus, com entidades fabricadas, é imune à primeira ressalva — descrever o inexistente é confabulação por definição — e constitui a evidência mais robusta deste trabalho.

6 A IA COMO FERRAMENTA DE APOIO: O MARCO REGULATÓRIO (CFM Nº 2.454/2026)

A Resolução CFM nº 2.454, de 11 de fevereiro de 2026, que normatiza o uso da inteligência artificial na medicina, fixa um princípio inegociável: a IA é instrumento de apoio, e a decisão permanece humana. O médico deve *"empregar a IA exclusivamente como ferramenta de apoio, mantendo-se como responsável final pelas decisões clínicas, diagnósticas, terapêuticas e prognósticas"* (art. 4º, I), exercendo julgamento crítico sobre as recomendações fornecidas (art. 4º, II). Reforçando-o, o art. 18 determina que *"em nenhum momento"* os sistemas de IA poderão *"restringir ou substituir a autoridade final do médico"*, e que a decisão *"caberá sempre ao médico, que pode acolher ou rejeitar as sugestões da IA conforme seu julgamento"*; o art. 15, parágrafo único, é categórico: as soluções de IA *"não são soberanas, sendo obrigatória a supervisão humana"* (CONSELHO FEDERAL DE MEDICINA, 2026).

Esse desenho normativo tem consequência prática direta para o presente trabalho. Se a IA não pode decidir pelo profissional — e se este responde criticamente por cada recomendação —, então cabe ao profissional aprofundar-se na *customização* e na *curadoria* dos conteúdos que alimentam a ferramenta, de modo que as respostas geradas fiquem aderentes às práticas e à evidência por ele reconhecidas. A própria Resolução aponta nessa direção ao recomendar, no Anexo III, a adoção preferencial de soluções que permitam *"customização e adaptação ao contexto local"*, dando preferência às que ofereçam *"possibilidade de treinamento adicional com dados locais e interfaces auditáveis, em detrimento de sistemas totalmente fechados"* (CONSELHO FEDERAL DE MEDICINA, 2026). Uma plataforma curada com conhecimento validado por um profissional experiente — como o NTAssist sobre a arquitetura *Cérebro do Mentor* — atende precisamente a esse requisito e, como demonstrado nas Seções 4 e 5, excede em muito o desempenho de uma IA generalista no tema curado.

Há, ainda, convergência entre as métricas deste benchmark e os atributos exigidos pela norma. A rastreabilidade e a fidelidade ao acervo medem, em termos operacionais, a *explicabilidade*, a *auditabilidade* e a *contestabilidade* das respostas — propriedades que a Resolução erige em condição de uso ético (Anexo I). Um sistema cujas afirmações são atribuíveis a fontes citáveis é, por construção, auditável pelo médico presente; um modelo generalista que confabula afirmações não-rastreáveis frustra a supervisão humana que a norma torna obrigatória. Do

mesmo modo, o direito do médico de recusar sistemas *"que não apresentem validação científica adequada"* (art. 3º, III) e o dever de transparência por *"indicadores científicos comprobatórios da acurácia, eficácia e segurança"* (art. 1º, § 3º) encontram, em avaliações como a aqui reportada, o tipo de evidência que a regulação pressupõe (CONSELHO FEDERAL DE MEDICINA, 2026).

7 CURADORIA E ADESÃO À PRÁTICA: A RECUSA DA "McDONALDIZAÇÃO" DO CUIDADO

A exigência regulatória de centralidade humana ecoa uma preocupação mais antiga com a desumanização da medicina. Dorsey e Ritzer (2016) descreveram a *"McDonaldização"* da medicina como a dominação do cuidado por quatro princípios da lógica do fast-food — eficiência, calculabilidade, previsibilidade e controle por tecnologias não-humanas —, cujo excesso *"nega a humanidade [...] das pessoas servidas"* por esses sistemas e ameaça o cuidado do indivíduo e as relações significativas entre médico e paciente. O risco que os autores identificam na padronização — a homogeneização irracional do cuidado, antitética a uma assistência responsiva às preferências, necessidades e valores de cada paciente — é exatamente o que uma IA generalista, treinada para a resposta média e mais provável, tende a amplificar.

É aqui que a curadoria por especialista inverte o vetor. Em vez de homogeneizar, ela ancora as respostas na prática específica e validada de um profissional experiente, preservando a singularidade do acervo e, por extensão, do raciocínio que dele decorre. A "McDonaldização" da saúde — uma linha de produção em que o profissional já não pensa a vida do paciente — não pode ocorrer quando a ferramenta é desenhada para servir ao julgamento do profissional, e não para substituí-lo. Curar conteúdos, analisar os casos correlatos e construir soluções como o NTAssist são, nesse sentido, atos de resistência a essa lógica: devolvem ao profissional o papel de quem decide o que deve ser exposto e como.

A literatura recente reforça que essa devolução é, paradoxalmente, o que a boa IA viabiliza. Martinelli *et al.* (2026) argumentam que a tecnologia não encerra a figura do médico, mas remove a carga administrativa que o havia afastado do leito — em prática ambulatorial, profissionais despendem 49% do dia em registros eletrônicos e trabalho de mesa, contra apenas 27% em contato clínico direto (SINSKY *et al.*, 2016). Ao absorver tarefas substituíveis (documentação, codificação, triagem), a IA *"não substitui o julgamento do médico, ela o preserva"*; e, crucialmente, *"um médico ausente não pode auditar um algoritmo, mas um presente, sim"* (MARTINELLI *et al.*, 2026). O ganho não é a velocidade, mas o tempo

reconquistado para a escuta qualificada e a empatia que a própria Resolução protege (art. 5º) — e para compreender a história de vida de cada paciente, condição de um diagnóstico mais preciso.

8 O QUE OS CASOS ENSINAM: CONFABULAÇÃO, ANCORAGEM E COGNIÇÃO PRESERVADA

Casos documentados ilustram, fora do ambiente controlado deste benchmark, tanto o risco da confabulação quanto o valor da ancoragem. Khullar (2025) relata o episódio de um paciente que, ao perguntar a um chatbot generalista por substitutos do sal de cozinha, recebeu a sugestão de brometo — substância tóxica — e desenvolveu quadro neuropsiquiátrico grave; ao replicarem a consulta, o modelo repetiu a recomendação. O mesmo autor descreve como um modelo de diagnóstico baseado em recuperação (RAG), alimentado com a descrição não-estruturada de um caso, *"fabricou valores laboratoriais, sinais vitais e achados de exame"* inexistentes e chegou a um diagnóstico equivocado; quando recebeu os mesmos dados ordenados por saliência clínica, ancorou-se na evidência pertinente e acertou. A lição é dupla: a confabulação é endêmica aos generalistas, e a qualidade do que se recupera e se cura determina a qualidade da resposta — precisamente o mecanismo do NTAssist.

Não por acaso, plataformas que adotam ancoragem explícita seguem o mesmo princípio do acervo curado: a ferramenta OpenEvidence *"cita um conjunto de artigos revisados por pares, por vezes incluindo uma figura exata ou uma citação textual [...], para prevenir alucinações"* e, diante de um caso, não tenta de imediato resolver o mistério, mas formula perguntas de esclarecimento (KHULLAR, 2025). É a mesma postura epistêmica que o benchmark mediu como *abstenção honesta*: reconhecer o limite antes de afirmar.

Os casos também advertem contra a delegação acrítica. Estudos citados por Khullar (2025) mostram que o GPT-4 respondeu incorretamente cerca de dois terços de perguntas médicas abertas e que a fração de respostas com ressalvas ("não sou qualificado para dar conselho médico") caiu de mais de um quarto, em 2022, para apenas 1% — exatamente a perda de honestidade epistêmica que a curadoria e a abstenção combatem. Adverte-se, ainda, para o *descalçamento cognitivo* (*cognitive de-skilling*): gastroenterologistas que passaram a depender de IA para detectar pólipos tornaram-se piores em encontrá-los sem o auxílio. Daí a fórmula que reconcilia eficácia e responsabilidade — *"treinar médicos que saibam usar a IA, mas também saibam pensar"* — e a visão de Dhaliwal, para quem o melhor uso clínico da IA é o de orientação de percurso (*wayfinding*): identificar tendências na trajetória do paciente e

detalhes que escaparam, não ditar a conduta (KHULLAR, 2025). Iniciativas de campo confirmam o ganho quando a IA opera como apoio supervisionado: na rede Penda Health (Quênia), clínicos que usaram um assistente de IA em segundo plano cometeram 16% menos erros diagnósticos e 13% menos erros de tratamento (KHULLAR, 2025).

Em síntese, os casos convergem para a tese deste artigo: a IA generalista, livre de ancoragem, confabula e pode causar dano; a IA curada por especialista, ancorada e auditável, transforma-se em ferramenta legítima de apoio. A diferença não é apenas técnica — é a diferença entre uma resposta que *parece* certa e uma resposta *rastreável*, sob a palavra final do profissional.

9 CONSIDERAÇÕES FINAIS

Quando o caso de uso exige respostas fundamentadas, rastreáveis e honestas sobre um material curado por especialista, o NTAssist mostrou-se dramaticamente superior aos modelos generalistas: cerca de 4–9× mais fiel, 8–19× mais rastreável e 39–57× menos alucinador dentro do corpus, e — crucialmente — capaz de reconhecer os limites de sua base, abstendo-se em 100% das perguntas fora do corpus. A arquitetura *Cérebro do Mentor* demonstra que a combinação de curadoria por especialista, recuperação ancorada, composição restrita à evidência e abstenção explícita transforma um LLM de uso geral em um assistente de domínio confiável. Em contextos clínicos, em que a confabulação convincente é um risco e não apenas um erro, essa honestidade epistêmica é o diferencial decisivo.

Mais do que um ganho de métricas, o que se defende é um arranjo: a IA como ferramenta de apoio e o profissional de saúde com a palavra final sobre o conhecimento que deve ser exposto, na exata medida em que a Resolução CFM nº 2.454/2026 exige supervisão humana e veda a soberania do algoritmo. A curadoria por especialista é o que torna esse arranjo viável — alinha a ferramenta à prática do profissional, resiste à "McDonaldização" do cuidado e, ao automatizar o que é substituível, devolve-lhe tempo para o que é insubstituível: ouvir e compreender a história de vida de cada paciente. Nessa filosofia, o trabalho de curadoria, a análise dos casos correlatos e soluções como o NTAssist não competem com o profissional — existem para que ele volte a exercer, com apoio e segurança, a razão pela qual escolheu cuidar. Trabalhos futuros incluem validação com juízes humanos especialistas e a extensão do protocolo a outros domínios curados.

REFERÊNCIAS

BENDER, E. M. *et al.* On the dangers of stochastic parrots: can language models be too big? *In:* PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT). New York: ACM, 2021. p. 610–623.

CONSELHO FEDERAL DE MEDICINA (Brasil). Resolução CFM nº 2.454, de 11 de fevereiro de 2026. Normatiza o uso da inteligência artificial na medicina. *Diário Oficial da União*: seção 1, Brasília, DF, ed. 39, p. 158, 27 fev. 2026.

DORSEY, E. R.; RITZER, G. The McDonaldization of medicine. *JAMA Neurology*, Chicago, v. 73, n. 1, p. 15–16, 2016. DOI 10.1001/jamaneurol.2015.3449.

ES, S. *et al.* RAGAS: automated evaluation of retrieval augmented generation. *In:* PROCEEDINGS OF THE 18TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (EACL): System Demonstrations. Stroudsburg: ACL, 2024. p. 150–158.

GAO, Y. *et al.* Retrieval-augmented generation for large language models: a survey. *arXiv preprint* arXiv:2312.10997, 2023.

HUANG, L. *et al.* A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, New York, v. 43, n. 2, p. 1–55, 2025.

JI, Z. *et al.* Survey of hallucination in natural language generation. *ACM Computing Surveys*, New York, v. 55, n. 12, p. 1–38, 2023.

KADAVATH, S. *et al.* Language models (mostly) know what they know. *arXiv preprint* arXiv:2207.05221, 2022.

KHULLAR, D. If A.I. can diagnose patients, what are doctors for? *The New Yorker*, New York, 22 set. 2025. Disponível em: https://www.newyorker.com/magazine/2025/09/29/if-ai-can-diagnose-patients-what-are-doctors-for. Acesso em: 22 jun. 2026.

LEWIS, P. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. *In:* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS), v. 33. Red Hook: Curran Associates, 2020. p. 9459–9474.

MARTINELLI, C. *et al.* Artificial intelligence is not the end of the physician. *JAMA*, Chicago, 2026. Publicação eletrônica antecipada. DOI 10.1001/jama.2026.4356.

MAYNEZ, J. *et al.* On faithfulness and factuality in abstractive summarization. *In:* PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL). Stroudsburg: ACL, 2020. p. 1906–1919.

PAN, S. *et al.* Unifying large language models and knowledge graphs: a roadmap. *IEEE Transactions on Knowledge and Data Engineering*, v. 36, n. 7, p. 3580–3599, 2024.

PETRONI, F. *et al.* Language models as knowledge bases? *In:* PROCEEDINGS OF THE 2019 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP-IJCNLP). Stroudsburg: ACL, 2019. p. 2463–2473.

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 (ACL). Stroudsburg: ACL, 2018. p. 784–789.

RITZER, G. *The McDonaldization of society*. 7. ed. Thousand Oaks: Sage, 2013.

SHUSTER, K. *et al.* Retrieval augmentation reduces hallucination in conversation. *In:* FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: EMNLP 2021. Stroudsburg: ACL, 2021. p. 3784–3803.

SINSKY, C. *et al.* Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annals of Internal Medicine*, Philadelphia, v. 165, n. 11, p. 753–760, 2016. DOI 10.7326/M16-0961.

TUKEY, J. W. Comparing individual means in the analysis of variance. *Biometrics*, v. 5, n. 2, p. 99–114, 1949.

ZHENG, L. *et al.* Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *In:* ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS), v. 36. Red Hook: Curran Associates, 2023. p. 46595–46623.