26º ENANCIB
Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
A Amazônia na sociedade global em rede: assimetrias e potencialidades informacionais, tecnológicas e ambientais
Belém, Pará, 09 a 13 de novembro de 2026
Universidade Federal do Pará

ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO
ISSN 2177-3688

GT 11 – Informação & Saúde

Victor Felipe de Lima – Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP) - Campus de Marília
José Eduardo Santarem Segundo – Universidade de São Paulo (USP) / Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP) - Campus de Marília

PEQUENOS MODELOS DE LINGUAGEM: POSSIBILIDADES PARA SAÚDE COLETIVA NA AMAZÔNIA

SMALL LANGUAGE MODELS: *OPPORTUNITIES FOR COLLECTIVE HEALTH IN THE AMAZON*

Modalidade: Trabalho Completo

Resumo: A região Amazônica brasileira apresenta profundas desigualdades de conectividade que, em suas áreas isoladas, impossibilitam o uso de ferramentas de inteligência artificial baseadas em nuvem e de redes estáveis de telemedicina. Profissionais de saúde e agentes comunitários que atuam nesses territórios, contudo, portam dispositivos móveis pessoais. Este trabalho propõe e implementa um arranjo informacional e tecnológico fundamentado em pequenos modelos de linguagem, executados de modo integralmente desconectado nesses dispositivos, com geração aumentada por recuperação apoiada em manuais oficiais do Sistema Único de Saúde. O arranjo articula a curadoria documental organizada em 6 pilares temáticos em uma base compacta de borda; a produção de um conjunto de dados sintético composto por mil diálogos clínicos em português brasileiro, com termos regionais e 18,5% de exemplos de recusa; o ajuste fino quantizado do modelo-base; e um protocolo de avaliação centrado na qualidade da informação. A avaliação, conduzida com conjunto inédito de 33 itens, registrou taxa de alucinação de dosagem de 3%, entendendo-se por alucinação de dosagem a ocorrência de dosagens numéricas ausentes do contexto recuperado. Conclui-se, de forma preliminar, que a curadoria, a organização e a indexação da informação se mostraram mais determinantes para a confiabilidade do arranjo do que o porte do modelo, o que sugere um caminho para mitigar assimetrias informacionais sem depender de infraestruturas globais de satélite.

Palavras-chave: tecnologia da informação; recuperação da informação; inteligência artificial; inclusão digital; saúde coletiva.

Abstract: The Brazilian Amazon Region exhibits profound connectivity inequalities that, in its isolated areas, preclude the use of cloud-based artificial intelligence tools and stable telemedicine networks. However, health professionals and community health workers operating in these territories carry personal mobile devices. This study proposes and implements an informational and technological framework based on small language models, executed entirely offline on these devices, with retrieval-augmented generation supported by official manuals from the Brazilian Unified Health System. The framework articulates documentary curation organized into six thematic pillars within a compact edge database; the production of a synthetic dataset comprising one thousand clinical dialogues in Brazilia Portuguese, incorporating regional terminology and 18,5% refusal examples; the quantized fine-tuning of the base model; and an evaluation protocol centered on information quality. The evaluation, conducted with an unseen set of 33 items, recorded a dosage-hallucination rate of 3%, understood as the occurrence of numeric dosages absent from the retrieved context. The results preliminarily indicate that, in the experiments conducted, the curation, organization, and indexing of information was more determinant of system reliability than model scale, suggesting a path to mitigate informational asymmetries without dependence on global satellite infrastructures.
Keywords: information technology; information retrieval; artificial intelligence; digital inclusion; collective health.

1 INTRODUÇÃO

A Amazônia brasileira abriga comunidades ribeirinhas, quilombolas e indígenas atravessadas por barreiras geográficas, assistenciais e comunicacionais que dificultam o acesso regular a serviços especializados de saúde. Esse cenário expressa desigualdades digitais e informacionais, nas quais a inclusão ou exclusão de redes condiciona oportunidades de acesso, uso e apropriação da informação (Castells, 1999; Van Dijk, 2005). Nesses territórios, a tomada de decisão clínica é afetada pelas grandes distâncias, pela sazonalidade dos rios e pela escassez de profissionais especializados (Taveira; Scherer; Diehl, 2014). Ao mesmo tempo, soluções digitais e de inteligência artificial (IA) em saúde dependem de infraestrutura técnica adequada, incluindo eletricidade, conectividade à Internet, redes móveis e dispositivos, condições que permanecem desigualmente distribuídas e podem limitar seu uso em contextos de maior vulnerabilidade (Organização Mundial da Saúde, 2021). Sob a perspectiva da Ciência da Informação (CI), essas barreiras podem ser compreendidas como limitações ao acesso, ao uso e à comunicação efetiva da informação, reforçando assimetrias informacionais (Saracevic, 1999) e, no campo da saúde, contribuindo para a ampliação de desigualdades assistenciais em territórios já vulnerabilizados.

Na perspectiva da CI, o desafio não está na ausência de conhecimento clínico, já sistematizado em documentos oficiais do Ministério da Saúde, mas em sua curadoria, organização, recuperação e disponibilização local. Nesse sentido, a noção de "informação como coisa" de Buckland (1991) oferece base para compreender a documentação oficial como recurso informacional passível de ser estruturado para apoiar decisões em ambientes sem conectividade.

Diante da ampla disponibilidade de *smartphones* entre profissionais de saúde e agentes comunitários, este trabalho desloca o foco da expansão da infraestrutura de rede para a curadoria do conhecimento, alinhando-se ao debate sobre as assimetrias e potencialidades informacionais da Amazônia na sociedade em rede (Castells, 1999).

Defende-se a tese de que a confiabilidade de sistemas de IA em saúde depende mais da curadoria, organização e recuperação da informação do que do porte dos grandes modelos de linguagem[1] (*Large Language Models* – LLMs). Embora técnicas como ajuste fino quantizado (Dettmers; Lewis; Shleifer; Zettlemoyer, 2022) e geração aumentada por recuperação (*Retrieval-Augmented Generation* – RAG) (Lewis *et al.*, 2020) permitam especializar modelos compactos, é a fundamentação em bases documentais curadas que reduz alucinações – entendidas aqui como conteúdo gerado sem lastro nas fontes recuperadas – e assegura maior segurança às respostas (Ji *et al.*, 2023).

A pesquisa investiga como modelos de linguagem de pequeno porte, executados *offline* em dispositivos móveis e apoiados por RAG fundamentada em fontes oficiais, podem mitigar assimetrias informacionais em saúde na Amazônia isolada. Como contribuições, o estudo apresenta: (1) uma arquitetura híbrida composta por modelo de pequeno porte, recuperação local e ajuste fino quantizado, operando integralmente *offline*; (2) a curadoria de seis pilares documentais oficiais em uma base compacta de borda; (3) um conjunto de dados sintético com comportamento explícito de recusa; e (4) evidências empíricas preliminares que indicam a curadoria da informação como fator central de confiabilidade.

2 REFERENCIAL TEÓRICO

Nessa seção, apresentam-se os fundamentos que sustentam a articulação entre CI, modelos de linguagem, recuperação aumentada e computação no contexto da saúde.

[1] Modelos de linguagem treinados com grandes volumes de dados textuais, o que lhes confere a capacidade de compreender o texto, interpretar intenções e nuances e gerar respostas coerentes.

2.1 CIÊNCIA DA INFORMAÇÃO, ASSIMETRIA E INFORMAÇÃO COMO COISA

Na CI, o acesso, a recuperação e o uso da informação são centrais para compreender processos informacionais e suas implicações sociais (Saracevic, 1999). Em contextos de saúde, assimetrias informacionais tornam-se particularmente sensíveis, pois a ausência de informação no ponto de decisão pode comprometer a conduta clínica. Nesse sentido, a distinção de Buckland (1991) entre "informação como coisa", "informação como processo" e "informação como conhecimento" fundamenta a compreensão dos manuais oficiais, registros e bases documentais como recursos informacionais organizáveis, recuperáveis e mobilizáveis na ponta do cuidado.

No contexto de sistemas baseados em LLMs com RAG, essa compreensão documental ganha uma dimensão operacional: a seleção, a estruturação e a organização dos documentos que compõem o corpus deixam de ser etapas acessórias e passam a constituir condições centrais para a qualidade da recuperação e da geração. Nessa direção, Santarem Segundo (2026) aproxima a construção de corpus, a gestão de coleções, a organização da informação e a recuperação documental das práticas da CI, reforçando que a qualidade das respostas em sistemas RAG depende diretamente da curadoria e da gestão da base informacional utilizada.

Essa perspectiva também dialoga com os estudos de comportamento e uso da informação. Wilson (1999) situa a necessidade informacional como elemento central dos processos de busca; Belkin (1980), pela noção de estado anômalo de conhecimento, permite compreender situações em que o sujeito reconhece uma lacuna cognitiva; e Dervin (1998), com a abordagem *sense-making*, enfatiza a apropriação contextual da informação. Neste trabalho, esses referenciais sustentam a recuperação de trechos documentais contextualizados, em vez da dependência exclusiva de respostas memorizadas pelo modelo. Assim, em sistemas generativos de saúde, a questão principal desloca-se de quanto o modelo "sabe" para quão adequadamente a informação pertinente, rastreável e procedente chega ao momento da decisão (Buckland, 1991; Saracevic, 1999).

2.2 PEQUENOS MODELOS DE LINGUAGEM, AJUSTE FINO E GERAÇÃO AUMENTADA POR RECUPERAÇÃO

Na IA generativa, o debate recente não se restringe ao aumento de escala, embora as leis de escala indiquem ganhos associados a parâmetros, dados e capacidade computacional

(Kaplan *et al.*, 2020). Os pequenos modelos de linguagem[2] (*Small Language Models* – SLMs), quando especializados com dados de qualidade, podem alcançar desempenho competitivo em tarefas delimitadas, como indicam a "família" *Phi* (Abdin *et al.*, 2024) e os modelos LLaMA (Touvron *et al.*, 2023). Esse movimento é viabilizado por técnicas de ajuste fino eficiente e compressão. A adaptação de baixo posto (*Low-Rank Adaptation* – LoRA) reduz o número de parâmetros treináveis (Hu *et al.*, 2022), enquanto a sua versão quantizada (*Quantized LoRA* – QLoRA) combina adaptação de baixa dimensão e quantização em 4 *bits*, diminuindo o custo de treinamento (Dettmers; Pagnoni; Holtzman; Zettlemoyer, 2023). Métodos como *GPTQ* (*Generalized Post-Training Quantization*) também reduzem o custo de inferência com perda controlada de desempenho (Frantar; Singh; Alistarh, 2024), favorecendo modelos compactos, especializados e executáveis localmente em dispositivos móveis.

A RAG combina modelos generativos com mecanismos de recuperação externa, permitindo que as respostas sejam fundamentadas em trechos de uma base controlada, e não apenas no conhecimento paramétrico do modelo (Lewis *et al.*, 2020). Revisões recentes indicam que a qualidade da base, da recuperação e da procedência das fontes influencia diretamente a fidelidade das respostas (Gao *et al.*, 2024). Sob a perspectiva da CI, Santarem Segundo (2026) reforça que extração, segmentação em *chunks* (blocos de texto), *embeddings* (representações vetoriais densas do conteúdo), bancos de dados vetoriais, busca e composição do *prompt* (instrução/texto de entrada) dependem de processos de organização, representação e recuperação da informação.

Do ponto de vista técnico, a recuperação pode combinar métodos esparsos, como BM25 (*Best Matching 25*) (Robertson; Zaragoza, 2009), e métodos densos, como DPR (*Dense Passage Retrieval*) e *Sentence-BERT*, capazes de capturar relações semânticas além da coincidência lexical (Karpukhin *et al.*, 2020; Reimers; Gurevych, 2019). Índices vetoriais, como FAISS (*Facebook AI Similarity Search*), favorecem buscas aproximadas eficientes em grandes coleções (Johnson; Douze; Jégou, 2021). Em saúde, essa arquitetura é relevante porque a pergunta do profissional e o conteúdo do manual podem empregar vocabulários distintos. Assim, a combinação entre fonte oficial, recuperação local e rastreabilidade documental contribui para reduzir riscos de alucinação e ampliar a segurança informacional (Ji *et al.*, 2023).

[2] Modelos de linguagem compactos, com parâmetros e volume de dados de treinamento reduzidos em comparação com os LLMs, sendo mais otimizados e rápidos.

2.3 INTELIGÊNCIA ARTIFICIAL NA SAÚDE E EM DISPOSITIVOS DE BORDA

Na saúde, a IA tem sido associada ao apoio à decisão, à triagem, à sumarização e à ampliação da capacidade analítica dos serviços (Topol, 2019). Em sistemas generativos, contudo, a fluência textual não é suficiente: avaliações devem considerar fidelidade ao contexto, precisão factual e capacidade de recusa diante de informação ausente ou insuficiente (Es; James; Anke; Schockaert, 2024; Min *et al.*, 2023). A Organização Mundial da Saúde (2021) também destaca princípios de governança ética para IA em saúde, como transparência, responsabilidade e proteção de dados.

Nesse contexto, a execução em dispositivos de borda[3] torna-se relevante por reduzir dependência de nuvem, latência e exposição de dados, especialmente em territórios com conectividade instável. No Brasil, a telessaúde aplicada à saúde indígena e a atenção a populações remotas já constituem temas relevantes, mas ainda dependem amplamente de conexão à rede (Taveira; Scherer; Diehl, 2014). Essa dependência evidencia a necessidade de soluções que funcionem *offline*, no próprio dispositivo utilizado pelo profissional.

2.4 LACUNA E POSICIONAMENTO DO TRABALHO

Apesar dos avanços em modelos compactos, RAG, IA em saúde e computação de borda, ainda há pouca articulação empírica entre conhecimento oficial do Sistema Único de Saúde (SUS), indexação vetorial, pequenos modelos de linguagem e operação *offline* na ponta do cuidado amazônico. A abordagem desse problema pela CI, para a qual organização, recuperação, uso, relevância e qualidade da informação são variáveis centrais, e não apenas aspectos técnicos da implementação, permanece ainda mais incipiente (Buckland, 1991; Saracevic, 1999).

A literatura sobre RAG enfatiza a importância da base recuperada para a qualidade das respostas, mas se concentra majoritariamente em contextos conectados e em bases amplamente disponíveis, sobretudo em língua inglesa (Gao *et al.*, 2024; Lewis *et al.*, 2020). Já a literatura sobre telessaúde e saúde amazônica documenta desafios assistenciais em territórios remotos (Taveira; Scherer; Diehl, 2014), mas ainda apresenta poucas soluções de inteligência artificial operando integralmente *offline* na ponta do cuidado. É nessa interseção entre SLMs,

[3] Instrumentos inteligentes, como *smartphones* ou *wearables* (dispositivos vestíveis, como pulseiras ou relógios inteligentes) que processam informações no próprio local onde elas são criadas, em vez de enviar tudo para um servidor distante na nuvem.

recuperação local, fontes oficiais e dispositivos móveis desconectados, analisada sob a ótica informacional, que este trabalho se posiciona.

3 PROCEDIMENTOS METODOLÓGICOS

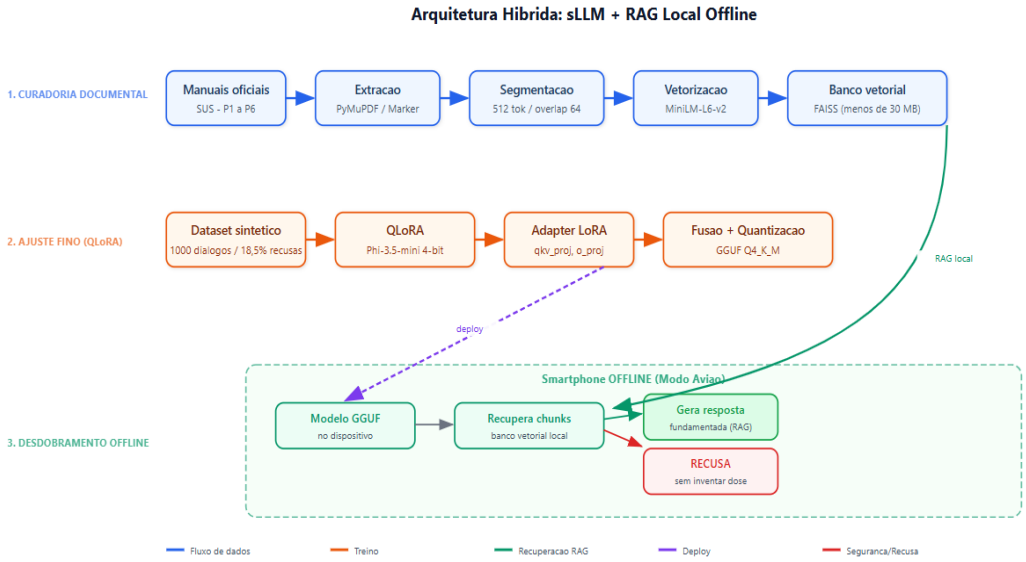
Trata-se de uma pesquisa aplicada, de natureza quali-quantitativa, do tipo exploratória, de abordagem empírica e orientada ao desenvolvimento e à avaliação de um artefato informacional-tecnológico. O percurso metodológico organiza-se em quatro frentes articuladas: curadoria documental de fontes oficiais; construção e curadoria de um conjunto de dados sintético para ajuste fino; ajuste fino quantizado e implantação *offline*; e avaliação centrada na qualidade da informação. O recorte territorial é a Amazônia isolada, enquanto o recorte temático foca na atenção primária e na saúde indígena, representadas pelos pilares 3 e 5 do corpus, enquanto os demais pilares compõem a base de conhecimento de borda para uso em inferência.

Os conceitos apresentados no referencial teórico foram operacionalizados nos procedimentos aqui descritos: a noção de informação como coisa (Buckland, 1991) orientou o tratamento dos manuais oficiais como artefatos documentais a serem selecionados, segmentados, indexados e registrados com procedência em cada bloco da base de borda; os estudos de comportamento informacional (Wilson, 1999) fundamentaram a definição dos pilares temáticos e dos cenários clínicos a partir das necessidades informacionais do território; o estado anômalo de conhecimento (Belkin, 1980) inspirou os cenários de recusa, nos quais o sistema deve reconhecer a lacuna informacional do contexto e abster-se de simular conhecimento; e a abordagem *sense-making* (Dervin, 1998) orientou a avaliação da adaptação linguística regional, tratada como condição de apropriação da informação pelo usuário.

3.1 ARRANJO GERAL E ARQUITETURA

A Figura 1 apresenta o arranjo geral proposto, organizado em quatro camadas:

Figura 1 – Arquitetura híbrida: modelo de pequeno porte com recuperação local *offline*



Fonte: elaborada pelos autores (2026).

Como apresentado na figura anterior, a primeira camada, voltada à curadoria, realiza a ingestão de documentos oficiais e sua preparação para recuperação. A segunda camada, de borda, mantém um banco vetorial compacto, capaz de operar diretamente no dispositivo. A terceira camada hospeda o SLM, submetido a ajuste fino e quantização. Por fim, a quarta camada realiza a implantação do sistema: o modelo fundido e quantizado é copiado para o smartphone, juntamente com o banco vetorial, e executado em modo avião.

A separação entre a base de conhecimento, atualizável independentemente do modelo, e o modelo generativo é uma decisão deliberada de arquitetura. Essa separação permite que a curadoria da informação, tratada neste trabalho como condição-chave de confiabilidade, seja revisada sem a necessidade de retreinar o modelo. A arquitetura híbrida apresentada combina um SLM com RAG para funcionar totalmente *offline* em um *smartphone*.

A cadeia inicia-se com a curadoria de manuais oficiais do SUS, que são extraídos, segmentados em blocos de 512 tokens, vetorizados com MiniLM-L6-v2 e armazenados em um banco vetorial FAISS de menos de 30 MB. Concomitantemente, um conjunto de dados sintético de 1.000 diálogos (18,5% deles de recusa) é usado para ajustar, via QLoRA, SLMs consagrados (Phi-3.5 e Qwen-0.5B, em 4 *bits*), gerando um adaptador LoRA que depois é fundido e quantizado em formato GGUF (*GPT-Generated Unified Format*).

Durante a execução, os modelos recuperam trechos relevantes do banco vetorial local e geram respostas fundamentadas nessas evidências. Em situações de risco, como a

solicitação de indicação de dose de medicamento, o sistema aciona uma recusa de segurança, em vez de arriscar uma resposta não fundamentada ou potencialmente inventada.

3.2 CURADORIA DOCUMENTAL E BASE DE CONHECIMENTO DE BORDA

O corpus é composto por seis pilares temáticos oficiais (Quadro 1), selecionados a partir das necessidades informacionais do território e das restrições de memória de dispositivos comuns. O *pipeline* de ingestão envolve extração de texto limpo, descarte de elementos não informacionais, segmentação em blocos de aproximadamente 512 *tokens* com sobreposição de 64 *tokens*, vetorização semântica em 384 dimensões e armazenamento em banco vetorial de borda.

Quadro 1 – Matriz de integração dos dados: pilares do corpus de borda

Pilar	Categoria temática	Origem oficial	Foco de extração
P1	Malária	Guia de Tratamento da Malária no Brasil	Esquemas de dose por idade; Tafenoquina; recaídas
P2	Animais peçonhentos	Manual de Acidentes por Animais Peçonhentos	Identificação de gêneros; ampolas de soro; complicações
P3	Saúde indígena	Política Nacional de Atenção à Saúde dos Povos Indígenas	Medicina tradicional; interculturalidade; fluxos
P4	Doenças negligenciadas	Manuais de Leishmaniose, Doença de Chagas, Hanseníase	Manejo; Doença de Chagas aguda oral via açaí
P5	Atenção primária	Cadernos de Atenção Básica	Condutas emergenciais; transporte fluvial
P6	Toxicologia	Manual de Populações Expostas a Mercúrio	Sinais neurológicos; garimpo; populações ribeirinhas

Fonte: elaborado pelos autores (2026).

A escolha do tamanho dos blocos equilibra duas exigências técnicas e informacionais: blocos muito pequenos tendem a perder contexto clínico, enquanto blocos muito grandes podem exceder a janela do modelo e diluir a relevância da recuperação. Cada bloco registra sua procedência, incluindo documento, edição e página, atributo central para auditoria, rastreabilidade e mitigação da desinformação. O volume compactado, com cerca de 360 mil *tokens*, ocupa menos de 30 MB, validando a densidade informacional.

3.3 CONJUNTO DE DADOS SINTÉTICO DE AJUSTE FINO

Para adaptar o modelo ao domínio linguístico, comunicacional e cultural da Amazônia isolada, foi construído um conjunto sintético de 1.000 diálogos clínicos em português

brasileiro, estruturados em mensagens de sistema, usuário e assistente. A geração foi orientada por cenários curados dos pilares 3 e 5, acompanhados de trechos de contexto que simulam a recuperação semântica em produção, incorporando termos regionais como *rabeta, bubuia, baque, pajé, igarapé, polo-base* e *remor*. O conjunto incluiu deliberadamente cenários de recusa, nos quais informações sensíveis, como dosagens ou número de ampolas de soro, não estavam presentes no contexto recuperado, exigindo que o modelo explicitasse a limitação informacional e orientasse a consulta ao protocolo oficial ou à regulação. Após a geração, realizou-se curadoria pós-geração para normalização ortográfica, consolidação de 13 termos canônicos e descarte de ruído lexical, com registro para fins de reprodutibilidade.

3.4 AJUSTE FINO QUANTIZADO E DESDOBRAMENTO OFFLINE

O modelo-base adotado possui aproximadamente 3.8 bilhões de parâmetros e janela de contexto longa, adequada ao processamento simultâneo de múltiplos trechos documentais. O ajuste fino empregou quantização em 4 *bits*, dupla quantização, cálculo em 16 bits e adaptadores de baixo posto, com posto 16 e alfa 32, aplicados aos módulos de atenção fundida do modelo. O treinamento utilizou ainda *gradient checkpointing* (pontos de verificação de gradiente) e um otimizador paginado de 8 *bits*, sendo executado por 3 épocas em unidade de processamento gráfico de uso gratuito.

Essas escolhas visam conciliar qualidade e viabilidade de borda: a quantização reduz drasticamente o tamanho do modelo; os adaptadores de baixo *ranking* permitem especializar o comportamento com baixo custo; o *gradient checkpointing* e o otimizador paginado controlam o consumo de memória. No desdobramento, o adaptador é fundido ao modelo-base, convertido para o formato quantizado apropriado à execução em celular (arquivo de cerca de 2,2 GB), copiado, junto ao banco vetorial, para o *smartphone* e executado em modo avião, com uso de memória limitado a 4 GB, em aplicativos de código aberto.

3.5 PROTOCOLO DE AVALIAÇÃO

A avaliação articula cinco perguntas: fundamentação das respostas no contexto recuperado; recusa correta quando a informação necessária está ausente; segurança clínica (não fabricar dosagens); adaptação linguística regional; e viabilidade de borda. Aplicou-se um conjunto de 33 itens inéditos, distintos dos cenários de treino, cobrindo fundamentação com

informação presente, recusa com informação ausente, compreensão de termos regionais, casos adversos (tentativas de induzir dosagens inexistentes) e interculturalidade.

As métricas principais foram: taxa de recusa correta (proporção de recusas adequadas quando a informação está ausente); taxa de falsa recusa (recusas indevidas quando a informação está presente); F1 da recusa; e taxa de alucinação de dosagem, medida pela proporção de respostas com dosagens numéricas não presentes no contexto, cuja meta é próxima de zero. Ressalta-se que o termo “alucinação” é empregado aqui de forma restrita, referindo-se especificamente à invenção de dosagens numéricas ausentes do contexto recuperado, e não à totalidade dos possíveis erros factuais de modelos generativos.

A detecção de recusa utilizou marcadores linguísticos, enquanto a detecção de dosagem empregou padrões numéricos. O modelo-base e o modelo ajustado foram comparados em geração determinística e com o mesmo prompt de sistema, de modo a isolar o efeito do ajuste fino. Trata-se de avaliação automatizada, de caráter preliminar, com amostra reduzida (n = 33) e detecção baseada em marcadores linguísticos e padrões numéricos, não constituindo validação clínica.

4 RESULTADOS E DISCUSSÃO

Essa seção apresenta os resultados da implementação proposta e discute seu desempenho, suas limitações e suas implicações para a saúde na Amazônia isolada.

4.1 RESULTADO: CARACTERIZAÇÃO DO CONJUNTO DE DADOS

Foram produzidos 1.000 exemplos (P3 = 490; P5 = 510, conforme o Quadro 1), derivados de 16 cenários, totalizando ≈ 405 mil tokens, com diálogos multiturno (média de 1,8 turno), dos quais 185 (18,5%) correspondem a recusas fundamentadas. Os termos regionais aparecem bem distribuídos (rabeta: 429; passar mal: 321; igarapé: 320; polo-base: 309; pajé: 246; baque: 188), evidenciando adaptação ao léxico do território.

4.2 RESULTADO: VIABILIDADE DO AJUSTE FINO

O ajuste fino foi validado de ponta a ponta em uma unidade de processamento gráfico de uso gratuito (*Google Colab*). O treinamento foi estável, sem instabilidade de precisão, com curva de perda decrescente registrada em tempo real em um painel de métricas, e o adaptador resultante foi publicado em repositório aberto (*HuggingFace*). A execução

completa, composta por três épocas sobre os 1.000 exemplos, foi concluída em, aproximadamente, 40 minutos, demonstrando viabilidade computacional sem custo financeiro. Esse aspecto está alinhado à mitigação de assimetrias tecnológicas, pois a reprodução do modelo não exige recursos computacionais fora do alcance de equipes de pesquisa ou de serviços públicos.

4.3 RESULTADO: AVALIAÇÃO (BASE VERSUS AJUSTADO), EXEMPLOS E LEITURA QUALITATIVA

A Tabela 1 sintetiza o desempenho no conjunto de 33 itens inéditos:

Tabela 1[4] – Resultados no conjunto inédito (n = 33)

Modelo SLMs*	Taxa de recusa correta	Taxa de falsa recusa	F1 (recusa)	Alucinação de dosagem
Phi-3.5-mini-finetuned	0,929	0,266	0,813	0,03
Qwen-0.5B-finetuned	0,857	0,053	0,889	0,03

Fonte: elaborada pelos autores (2026).

Os resultados indicam que o *Phi-3.5-finetuned* alcançou recusa correta em, aproximadamente, 93% dos casos sem dosagem disponível no contexto, registrando apenas 1 falso negativo e com taxa de alucinação de apenas 3%. O ajuste fino preservou essa segurança, reduziu recusas indevidas, elevou a medida F1 e produziu respostas mais consistentes do ponto de vista profissional. A comparação entre modelos de portes distintos revelou um achado central: tanto o *Phi-3.5*, com 3.8 bilhões de parâmetros, quanto o *Qwen-0.5B-finetuned*, com 500 milhões de parâmetros, apresentaram a mesma taxa de alucinação de dosagem, correspondente a 3%. Embora o modelo maior tenha demonstrado maior consistência na formulação das recusas, o controle do risco mais crítico, ou seja, a invenção de dosagens, foi equivalente em ambos. Esses resultados, obtidos em avaliação automatizada e com amostra reduzida, apontam que, nas condições avaliadas, a curadoria e fundamentação do contexto pesaram mais do que o porte do modelo para o controle do risco crítico examinado, a invenção de dosagens.

A inspeção qualitativa confirma o padrão observado. Em casos de fundamentação com informação presente, o modelo responde de modo ancorado no contexto, por exemplo, ao orientar o reconhecimento de gravidade a partir de sinais descritos no trecho recuperado,

[4] Nota: **avaliação automatizada em conjunto inédito (n = 33), com métricas delimitadas conforme a Seção 3.5.**

como o batimento de asa de narina em criança com pneumonia. Em casos de recusa, diante da solicitação de uma dosagem ausente, como a dose de um antitérmico para criança indígena, quando o trecho recuperado trata apenas da organização do cuidado, o modelo declara a limitação e encaminha ao protocolo oficial, sem inventar valores numéricos. Em casos de compreensão de termos regionais, o modelo interpreta corretamente expressões como “bubuiando” (no sentido de flutuar à deriva da correnteza), ou “deu um baque” (associado a mal-estar agudo no contexto clínico), integrando-as ao raciocínio.

4.4 DISCUSSÃO: A CURADORIA INFORMACIONAL COMO CHAVE DA CONFIABILIDADE

O achado mais relevante para o campo, nas condições deste experimento, converte-se aqui na tese central do trabalho: o modelo-base, sem qualquer ajuste fino, já apresenta taxa de recusa correta igual a ≈ 0,93 e taxa de alucinação de dosagem próxima de 3%. Em outras palavras, a maior parte do comportamento de segurança decorre do prompt de sistema e, sobretudo, da fundamentação no contexto curado (não da escala do modelo nem do ajuste fino). Isso sustenta, empiricamente e de forma preliminar, a hipótese de que o tratamento, a organização, a indexação vetorial de dados públicos e a curadoria documental pesam, nas condições avaliadas, mais do que essa variável para a confiabilidade de sistemas de IA em saúde. O ajuste fino agrega estilo, consistência e menor excesso de recusa, refinando, sem substituir, uma arquitetura cuja espinha dorsal informacional é a recuperação (RAG) sobre fontes oficiais. Como implicação prática, investir em curadoria e procedência do corpus tende a produzir maior ganho de segurança por unidade de esforço do que ampliar o modelo, ainda que essa leitura demande confirmação experimental dedicada, por meio de delineamentos que variem controladamente a qualidade da curadoria e tamanho dos modelos.

O comportamento seguro observado decorre menos do conhecimento paramétrico do modelo e mais da qualidade do artefato informacional disponibilizado no momento da decisão, expressão concreta da informação como coisa (Buckland, 1991): são os blocos curados, com procedência registrada, que condicionam a resposta gerada. Do ponto de vista do estado anômalo de conhecimento (Belkin, 1980), a recusa correta funciona como o reconhecimento explícito de que a lacuna informacional do usuário não pôde ser preenchida pelo corpus disponível, transferindo a decisão ao protocolo oficial (comportamento informacional desejável em saúde, e não falência do sistema). Na perspectiva do *sense-making* (Dervin, 1998), a adaptação a termos regionais reduz a distância entre a situação do usuário e

a informação oferecida, favorecendo sua apropriação. Os modelos de comportamento informacional (Wilson, 1999), por sua vez, ajudam a interpretar por que o mesmo sistema pode ser útil ou inócuo conforme o contexto de uso da informação pela equipe. Essa leitura é analítica, e não experimental, pois o delineamento aqui apresentado não isola o efeito da curadoria em relação ao porte do modelo.

Do ponto de vista da informação em saúde, recusas e alucinações têm implicações distintas e assimétricas. A alucinação de dosagem corresponde ao pior desfecho informacional possível: entrega conteúdo aparentemente válido, sem lastro no contexto recuperado, com potencial de dano assistencial direto; por isso sua meta é zero e sua ocorrência deve ser monitorada continuamente. A recusa correta, em contraste, sinaliza os limites do corpus disponível e redireciona o profissional ao protocolo oficial e à regulação – comportamento seguro do ponto de vista informacional, ainda que produza lentidão operacional, como o tempo adicional de consulta, e que, se excessivamente frequente, possa desgastar a confiança no sistema. Já a falsa recusa funciona como indicador de lacunas do corpus, onde cada falso negativo aponta conteúdo que deveria estar recuperável e não está, retroalimentando a curadoria documental. Em síntese, esses indicadores convertem-se em variáveis de gestão da informação e de segurança do paciente: a taxa de alucinação de dosagem mede o risco de dano; a taxa de falsa recusa mede a cobertura da base curada; e ambas dependem da manutenção, da atualização e da procedência do corpus, especialmente em contextos remotos, nos quais o sistema pode ser a única fonte imediata de orientação (Organização Mundial da Saúde, 2021; Taveira; Scherer; Diehl, 2014).

4.5 DISCUSSÃO: IMPLICAÇÕES PARA A CIÊNCIA DA INFORMAÇÃO

A contribuição desloca o foco analítico da escala do modelo para a qualidade da informação. Procedência, segmentação, indexação e governança do corpus passam a ser variáveis de primeira classe na confiabilidade do sistema, e não detalhes de engenharia. Esse deslocamento dialoga com a tradição da área, que há muito trata a mediação, a organização e a qualidade da informação como condições de uso confiável, e responde diretamente ao eixo temático: trata-se de uma potencialidade informacional da Amazônia, é possível capacitar o agente público na ponta do cuidado por meio de conhecimento de borda estruturado, mitigando assimetrias informacionais e tecnológicas sem depender de investimentos massivos

em infraestruturas globais de satélite. O núcleo da solução está na informação, sua procedência, sua organização e sua governança, e não na escala do modelo.

Os resultados dialogam com a literatura que associa a recuperação aumentada à redução de alucinações e que recomenda avaliação centrada em segurança, mais que em métricas de fluência. A novidade aqui é dupla: aplicar o princípio à borda desconectada, com fontes oficiais brasileiras, e lê-lo sob a ótica da CI.

5 CONSIDERAÇÕES FINAIS

Este estudo investigou como SLMs, capazes de operar offline em dispositivos móveis e baseados em fontes oficiais, podem mitigar desigualdades tecnológicas e informacionais na Amazônia sem depender de investimentos massivos em infraestrutura de satélite. Os experimentos, que combinaram curadoria documental oficial, um conjunto sintético de mil diálogos, ajuste fino quantizado dos modelos e um protocolo de avaliação voltado à qualidade da informação, mostraram que esse ajuste fino é viável em infraestrutura computacional gratuita, com treinamento prático, e que os modelos testados alcançaram elevadas taxas de recusa correta, boa precisão e baixa taxa de alucinação de dosagem, independentemente do tamanho do modelo. Assim, reforçando a hipótese de que a confiabilidade dos sistemas depende mais da curadoria e da organização da informação do que do porte do modelo, reafirmando, no campo da CI, a centralidade do tratamento e da curadoria de dados públicos para sistemas de IA seguros, principalmente em contextos de maior assimetria.

Dessa forma, os resultados demonstram viabilidade técnica e informacional promissora, mas que exige, por exemplo, de validação futura de especialistas, estudos de campo longitudinais e ampliação do conjunto de dados com situações reais. Como contribuições, o estudo oferece uma arquitetura híbrida reproduzível, um corpus curado em seis pilares temáticos, um conjunto sintético com recusa explícita e um protocolo de avaliação focalizado na qualidade da informação.

REFERÊNCIAS

ABDIN, Marah *et al*. **Phi-3 technical report**: a highly capable language model locally on your phone. Redmond: Microsoft, ago. 2024. Disponível em: https://www.microsoft.com/en-us/research/publication/phi-3-technical-report-a-highly-capable-language-model-locally-on-your-phone/. Acesso em: 7 jun. 2026.

BELKIN, Nicholas J. Anomalous states of knowledge as a basis for information retrieval. **Canadian Journal of Information Science**, [*s. l.*], v. 5, p. 133-143, 1980. Disponível em:

https://tefkos.comminfo.rutgers.edu/Courses/612/Articles/BelkinAnomolous.pdf. Acesso em: 9 jun. 2026.

BUCKLAND, Michael K. Information as thing. **Journal of the American Society for Information Science**, [*s. l.*], v. 42, n. 5, p. 351-360, jun. 1991. Disponível em: https://ppggoc.eci.ufmg.br/downloads/bibliografia/Buckland1991.pdf. Acesso em: 11 jun. 2026.

CASTELLS, Manuel. **A sociedade em rede**. São Paulo: Paz e Terra, 1999. v. 1. (A era da informação: economia, sociedade e cultura).

DERVIN, Brenda. Sense-making theory and practice: an overview of user interests in knowledge seeking and use. **Journal of Knowledge Management**, [*s. l.*], v. 2, n. 2, p. 36-46, 1998. DOI: https://doi.org/10.1108/13673279810249369. Acesso em: 10 jun. 2026.

DETTMERS, Tim; LEWIS, Mike; SHLEIFER, Sam; ZETTLEMOYER, Luke. 8-bit optimizers via block-wise quantization. *In*: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS, 10., 2022, Online. **Proceedings** [...]. [*S. l.*]: ICLR, 2022. p. 1-19. Disponível em: https://openreview.net/pdf?id=shpkpVXzo3h. Acesso em: 12 jun. 2026.

DETTMERS, Tim; PAGNONI, Artidoro; HOLTZMAN, Ari; ZETTLEMOYER, Luke. QLoRA: efficient finetuning of quantized LLMs. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 37., 2023, New Orleans. **Proceedings** [...]. New York: Curran Associates, 2023. p. 441. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf. Acesso em: 11 jun. 2026.

ES, Shahul; JAMES; Jithin; ANKE, Luis Espinosa; SCHOCKAERT, Steven. RAGAS: automated evaluation of retrieval augmented generation. *In*: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 18., 2024, St. Julian's. **Proceedings** [...]. St. Julian's: Association for Computational Linguistics, 2024. p. 150-158. DOI: https://doi.org/10.18653/v1/2024.eacl-demo.16. Acesso em: 9 jun. 2026.

FRANTAR, Elias; SINGH, Sidak; ALISTARH, Dan. Optimal brain compression: a framework for accurate post-training quantization and pruning. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 36., 2022, New Orleans. **Proceedings** [...]. New York: Curran Associates, 2022. p. 323. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2022/file/1caf09c9f4e6b0150b06a07e77f2710c-Paper-Conference.pdf. Acesso em: 9 jun. 2026.

GAO, Yunfan *et al*. Retrieval-augmented generation for large language models: a survey. **arXiv**, [S. l.], v. 5, p. 1-21, mar. 2024. Preprint. DOI: https://doi.org/10.48550/arXiv.2312.10997. Acesso em: 7 jun. 2026.

JI, Ziwei *et al*. Survey of hallucination in natural language generation. **ACM Computing Surveys**, New York, v. 55, n. 12, p. 1-38, mar. 2023. DOI: https://doi.org/10.1145/3571730. Acesso em: 6 jun. 2026.

JOHNSON, Jeff; DOUZE, Matthijs; JÉGOU, Hervé. Billion-scale similarity search with GPUs. **IEEE Transactions on Big Data**, [*s. l.*], v. 7, n. 3, p. 535-547, jul. 2021. DOI: https://doi.org/10.1109/TBDATA.2019.2921572. Acesso em: 8 jun. 2026.

KAPLAN, Jared *et al*. Scaling laws for neural language models. **arXiv**, [S. l.], p. 1-30, jan. 2020. Preprint. DOI: https://doi.org/10.48550/arXiv.2001.08361. Acesso em: 10 jun. 2026.

KARPUKHIN, Vladimir *et al*. Dense passage retrieval for open-domain question answering. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2020, Online. **Proceedings** [...]. [S. l.]: ACL, 2020. p. 6769-6781. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.550. Acesso em: 9 jun. 2026.

LEWIS, Patrick *et al*. Retrieval-augmented generation for knowledge-intensive NLP tasks. *In*: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 34., 2020, Vancouver. **Proceedings** [...]. New York: Curran Associates, 2020. p. 793. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf. Acesso em: 7 jun. 2026.

MIN, Sewon *et al*. FActScore: fine-grained atomic evaluation of factual precision in long form text generation. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2023, Singapore. **Proceedings** [...]. Singapore: ACL, 2023. p. 12076-12100. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.741. Acesso em: 9 jun. 2026.

ORGANIZAÇÃO MUNDIAL DA SAÚDE. **Ethics and governance of artificial intelligence for health**: WHO guidance. Geneva: World Health Organization, 2021.

REIMERS, Nils; GUREVYCH, Iryna. Sentence-BERT: sentence embeddings using Siamese BERT-networks. *In*: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND THE INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 9., 2019, Hong Kong. **Proceedings** [...]. Hong Kong: ACL, 2019. p. 3982-3992. DOI: https://doi.org/10.18653/v1/D19-1410. Acesso em: 11 jun. 2026.

ROBERTSON, Stephen; ZARAGOZA, Hugo. The probabilistic relevance framework: BM25 and beyond. **Foundations and Trends in Information Retrieval**, [*s. l.*], v. 3, n. 4, p. 333-389, set. 2009. Disponível em: https://www.researchgate.net/publication/220613776_The_Probabilistic_Relevance_Framework_BM25_and_Beyond. Acesso em: 8 jun. 2026.

SANTAREM SEGUNDO, José Eduardo. Modelos de linguagem (LLM) com geração aumentada por recuperação (RAG) na saúde: a Ciência da Informação no centro dos desafios e oportunidades. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 25., 2025, Rio de Janeiro. **Anais** [...]. Rio de Janeiro: ANCIB, 2026. Disponível em: https://enancib.ancib.org/enancib/pt_BR/article/view/1195. Acesso em: 21 jun. 2026.

SARACEVIC, Tefko. Information science. **Journal of the American Society for Information Science**, [*s. l.*], v. 50, n. 12, p. 1051-1063, out. 1999. Disponível em: https://tefkos.comminfo.rutgers.edu/JASIS1999.pdf. Acesso em: 8 jun. 2026.

TAVEIRA, Zaira; SCHERER, Magda; DIEHL, Eliana. Implantação da telessaúde na atenção à saúde indígena no Brasil. **Cadernos de Saúde Pública**, Rio de Janeiro, v. 30, n. 8, p. 1793-1797, ago. 2014. DOI: https://doi.org/10.1590/0102-311X00026214. Acesso em: 5 jun. 2026.

TOPOL, Eric. **Deep medicine**: how artificial intelligence can make healthcare human again. New York: Basic Books, 2019.

TOUVRON, Hugo *et al*. LLaMA: open and efficient foundation language models. **AI at Meta**, [*s. l.*], p. 1-27, fev. 2023. Disponível em: https://ai.meta.com/research/publications/llama-open-and-efficient-foundation-language-models/. Acesso em: 9 jun. 2026.

VAN DIJK, Jan. **The deepening divide**: inequality in the information society. Thousand Oaks: SAGE, 2005.

WILSON, Thomas D. Models in information behaviour research. **Journal of Documentation**, [*s. l.*], v. 55, n. 3, p. 249-270, jun. 1999. Disponível em: https://www.researchgate.net/publication/228784950_Models_in_Information_Behaviour_Research. Acesso em: 7 jun. 2026.