

MODELOS DE LINGUAGEM DE GRANDE ESCALA APLICADOS À SISTEMAS DE SAÚDE

CAMILA FANK KIST^{1,2*}, SAMUEL DA SILVA FEITOSA^{2,3}

1 Introdução

As feridas crônicas e agudas têm alta ocorrência e são um problema de saúde pública. Cuidar delas exige avaliar atributos como o tipo de tecido, a profundidade, a presença de infecção e o exsudato para, então, definir a conduta, inclusive quão urgente é o encaminhamento. Nem sempre as pessoas conseguem acesso imediato a especialistas. A inteligência artificial já vem sendo testada nessa avaliação e, apesar de desempenho comparável ao consenso de especialistas, modelos que analisam somente a imagem ainda falham em apresentações atípicas e cometem erros clinicamente graves em casos complexos (CORRADINI et al., 2025), e uma infecção tratada tarde demais pode evoluir para osteomielite, amputação e sepse (INTERNATIONAL WOUND INFECTION INSTITUTE, 2022). A dificuldade da tarefa é considerável: no benchmark WoundCareVQA, mesmo o GPT-4o fica entre 0,32 e 0,78 de acurácia por atributo, e a concordância entre os próprios especialistas varia de 0,32 a 0,92 (YIM et al., 2025). A execução local de modelos pequenos surge como alternativa, trazendo a vantagem de mitigar os custos de nuvem e preservar a privacidade. Esses modelos já foram aplicados nesse domínio, inclusive para decidir o encaminhamento (SAADATI FARD et al., 2025). Contudo, seu uso permanece fragmentado: seja como classificadores que, embora ofereçam explicabilidade via mapas de atribuição, não geram raciocínio em linguagem natural nem possuem verificação, seja restringindo-se a subtarefas isoladas, como a avaliação de infecções (BUSARANUVONG et al., 2026).

A partir dessa lacuna, este trabalho desenvolve e avalia o Wound Assessment Multi-Agent (WAMA), um pipeline modular que usa modelos pequenos (até 10 bilhões de parâmetros) rodando localmente. Ele é organizado em três estágios: um Perceptor, responsável pela percepção visual, um Residente, que raciocina sobre a conduta e o encaminhamento e Auditores, que verificam a saída. A contribuição não está em usar modelos pequenos, e sim em

1 Titulação acadêmica: Graduada em Ciência da Computação, instituição UFFS, *campus* Chapecó, contato: camila.kist@estudante.uffs.edu.br

2 Grupo de Pesquisa: IDT - Inovação e desenvolvimento tecnológico

3 Titulação acadêmica: Doutor, instituição UFFS, **Orientador(a)**.

integrá-los da percepção até a decisão de encaminhamento com uma camada de verificação e avaliá-los rigorosamente, usando F1-macro contra um piso ingênuo e o teto de concordância entre humanos, medindo também quanto o erro de percepção se propaga até a conduta.

2 Objetivos

Objetivo geral: desenvolver e avaliar um pipeline modular de modelos pequenos e locais para avaliação de feridas e decisão de encaminhamento. Objetivos específicos: (i) verificar se um modelo de visão-linguagem pequeno discrimina os atributos de uma ferida acima do piso ingênuo e próximo do teto de concordância humana; (ii) verificar se um modelo de linguagem pequeno transforma esse laudo em uma decisão de encaminhamento confiável; (iii) quantificar a propagação de erro entre a percepção e a conduta.

3 Metodologia

O pipeline se encadeia em três estágios. O Perceptor, um modelo de visão-linguagem, recebe a imagem e o relato e devolve um laudo estruturado com atributos. O Residente, um modelo de linguagem, parte desse laudo, do relato do paciente e de protocolos clínicos recuperados por RAG para produzir o raciocínio clínico, a conduta e o encaminhamento. Por fim, quatro Auditores independentes aprovam ou reprovam a saída do Residente.

A pesquisa utiliza o WoundCareVQA (YIM et al., 2025), no subconjunto de desenvolvimento (train.json, 279 casos), que reúne imagem, relato, sete atributos das imagens e a resposta em texto do especialista. Cabe uma ressalva sobre esse subconjunto: uma única pessoa o anotou, sem voto de maioria, o que compromete parte de sua confiabilidade.

A seleção dos modelos partiu de leaderboards, filtrando modelos de até 10B, com licença aberta e execução local. Depois, os selecionados foram testados no pipeline. Para o Perceptor, o modelo selecionado foi o InternVL3.5-8B, com desempenho superior ao do Qwen2.5-VL-7B e ao do MiniCPM-V-4.5. Para o Residente, foi adotado o nikita, um fine-tune médico do Qwen3.5-9B (o modelo de melhor desempenho entre os modelos pequenos avaliados), e descartado o MedGemma-4B e o Granite-4.1-8B por colapso técnico e de raciocínio, respectivamente.

Sobre as métricas: como as classes são muito desbalanceadas, foi adotado o F1-macro (média por classe) como métrica principal, e não a acurácia, que apenas premiaria o acerto na classe majoritária. O trabalho compara cada atributo com duas referências externas: um piso

ingênuo, que responde sempre a classe mais comum, e o teto de concordância entre anotadores (IAA) do WoundCareVQA. Nos atributos ordinais (com uma ordem natural entre categorias, como estágio ou intensidade de exsudato), acrescenta ainda um balde clínico: um agrupamento de categorias vizinhas em faixas clinicamente equivalentes, que tolera pequenos desvios de grau sem penalizá-los como erro. Já o Residente não possui um rótulo de conduta de referência para comparação direta. Por isso, sua avaliação foi feita sobre o eixo do encaminhamento: como o texto do especialista não indica explicitamente o grau de urgência, foi construído um proxy a partir de palavras-chave desse texto, numa escala de cinco níveis. Nessa avaliação, a sensibilidade mede a capacidade do modelo de captar a urgência real do caso, e a especificidade, sua capacidade de evitar encaminhamentos desnecessários. Uma camada determinística de reconciliação ajusta o rótulo de urgência gerado pelo modelo ao prazo de retorno que ele próprio descreveu no texto, corrigindo inconsistências internas da resposta.

Uma ressalva: estes resultados vêm do conjunto de desenvolvimento (train.json), sobre o qual o pipeline foi ajustado, e por isso refletem desempenho de desenvolvimento, não ainda de validação.

4 Resultados e Discussão

Perceptor: no cenário de referência (InternVL3.5-8B, temperatura 0, 279 casos e sem erro técnico), o modelo supera o piso ingênuo em F1-macro nos seis atributos de valor único, o que já responde à pergunta sobre discriminação. A Tabela 1 traz, por atributo, a acurácia e o F1-macro do modelo, os mesmos valores para o piso ingênuo, a faixa de concordância humana (teto IAA) e uma coluna de situação que classifica a acurácia do modelo como abaixo, dentro ou acima dessa faixa (Tabela 1).

Tabela 1 — Perceptor (InternVL3.5-8B, temperatura 0, n = 279).

Atributo	Acurácia	F1-macro	Piso: Acur. / F1	Teto IAA (acur.)	Situação
Tipo de ferida	0,70	0,37	0,86 / 0,13	0,865–0,892	abaixo
Espessura/estágio	0,36	0,24	0,36 / 0,09	0,324–0,405	no teto
Cor do tecido	0,56	0,33	0,49 / 0,11	0,568–0,622	abaixo
Qtd. de exsudato	0,57	0,24	0,55 / 0,12	0,514–0,541	acima*
Tipo de exsudato	0,65	0,46	0,57 / 0,15	0,649–0,676	no teto
Infecção	0,71	0,50	0,86 / 0,31	0,703–0,784	no teto

Piso: Acur./F1 = acurácia/F1-macro do piso ingênuo (classe majoritária). Teto IAA (acur.) = faixa de acurácia de concordância entre especialistas. Situação = acurácia do modelo vs. essa faixa (abaixo/no teto/acima*). O asterisco indica possível viés à classe majoritária em campo desbalanceado, não necessariamente desempenho melhor que o humano, por isso deve ser lido junto com o F1-macro. No balde clínico, a acurácia dos atributos ordinais sobe

para 0,69 (espessura), 0,75 (cor) e 0,74 (exsudato). Fonte: autoria própria, 2026.

O Qwen2.5-VL-7B tem acurácia maior, mas por tender à classe majoritária: seu F1-macro em infecção iguala o piso ingênuo e seu recall de infectados é zero, contra 0,41 do InternVL. Isso confirma que o F1-macro é a métrica clinicamente adequada aqui. Além disso, vários atributos já esbarram no teto da concordância humana, e também os próprios modelos base dos autores do WoundCareVQA chegam a esses mesmos valores. Notavelmente, nem os modelos grandes de nuvem (GPT-5.6, Gemini, Kimi) superam o piso ingênuo com folga. A tarefa resiste a ganhos de escala, reforçando o caminho dos modelos pequenos e locais.

Sobre o Residente: na rodada final (nikita, 279 casos), a confiabilidade técnica chegou a quase 100% (278 de 279 saídas válidas). Quando o especialista encaminha, o Residente também encaminha em 98% dos casos, que é o eixo clinicamente mais importante. A maior dificuldade está em determinar corretamente o grau de urgência: a sensibilidade sobe de 24% (com o LLM cru) para 37% depois da reconciliação determinística, ao preço de a especificidade cair de 94% para 87% (a camada não inventa urgência, apenas alinha o rótulo ao prazo que o próprio modelo escreveu). Descontando os cerca de sete casos em que o proxy do gabarito é impreciso, os números justos ficam em torno de 44% e 95%.

A respeito da propagação de erro, o Residente foi alimentado com o gabarito e com o laudo do Perceptor, e mudou a conduta em 26% dos casos. Na decisão de urgência, a percepção imperfeita quebrou 37 respostas corretas e consertou outras 23, com saldo líquido de -14 casos (a acurácia cai de 76% para 71%).

Os auditores foram testados de forma parcial (50 casos): apenas o verificador de conformidade protocolar reprova de forma relevante (10% a 16%), identificando tratamento excessivo (ex.: prata fora de protocolo). Os outros três aprovam quase tudo (96% a 100%), taxa que ainda exigirá validação por um avaliador humano.

5 Conclusão

Os resultados mostram que um pipeline modular de modelos pequenos e locais é viável e eficaz para o problema, discriminando os atributos visuais acima do piso ingênuo e decidindo o encaminhamento com segurança, sem os custos de nuvem e privacidade dos modelos grandes. Como a avaliação usou um conjunto de desenvolvimento com anotador único e algum ruído, esses valores tendem a ser um limite inferior do desempenho real. O ponto com maior margem de evolução é a decisão fina de urgência, e os próximos passos são validação em conjunto multi-anotador, validação clínica com um especialista e conclusão da camada de auditoria.

Referências Bibliográficas

BUSARANUVONG, P. et al. Infection-Reasoner: A Compact Vision-Language Model for Wound Infection Classification with Evidence-Grounded Clinical Reasoning. arXiv preprint arXiv:2604.19937, 2026.

CORRADINI, L. et al. AI vs. MD: Benchmarking ChatGPT and Gemini for Complex Wound Management. Journal of Clinical Medicine, v. 14, n. 24, art. 8825, 2025. DOI: 10.3390/jcm14248825.

INTERNATIONAL WOUND INFECTION INSTITUTE (IWII). Wound Infection in Clinical Practice: International Consensus Update 2022. London: Wounds International, 2022. Disponível em: <https://woundsinternational.com/wp-content/uploads/2023/05/IWII-CD-2022-web.pdf>.

SAADATI FARD, R. et al. Multimodal AI on Wound Images and Clinical Notes for Home Patient Referral. IEEE Journal of Translational Engineering in Health and Medicine, 2025. arXiv:2501.13247.

YIM, W. et al. WoundcareVQA: A multilingual visual question answering benchmark dataset for wound care. Journal of Biomedical Informatics, v. 170, art. 104888, 2025. DOI: 10.1016/j.jbi.2025.104888.

Palavras-chave: feridas crônicas; inteligência artificial; modelos de linguagem; triagem clínica; avaliação de feridas.

Nº de Registro no sistema Prisma: PES-2025-0314.

Financiamento: UFFS.

