

Médicos, Máquinas e Marketing Multimodal: Avaliando VLMs para Anotação Especializada de Anúncios Relacionados à Saúde

João F. H. Olivetti¹, Jônatas H. dos Santos², Marcos A. Gonçalves², Philipe de F. Melo¹
Julio C. S. Reis¹, Fabricio Benevenuto², Thales H. Silva², Katharina Borges²
Matheus Gontijo Guimarães², Glaucio de Souza², Marco A. G. Rodrigues², Cristiano X. Lima²
joao.olivetti@ufv.br, jonatashds@ufmg.br, mgoncalv@dcc.ufmg.br, philipe.freitas@ufv.br, jreis@ufv.br
fabricio@dcc.ufmg.br, thaleshenrique@ufmg.br, katharinasborges@gmail.com, guimaraesmatheus22@gmail.com
glauccio.gss@gmail.com, magro.mg@terra.com.br, cxlima@ufmg.br

¹Universidade Federal de Viçosa (UFV), Brasil

²Universidade Federal de Minas Gerais (UFMG), Brasil

ABSTRACT

Identifying medicalization in advertisements is a task that requires specialized knowledge and the joint interpretation of visual and textual elements. Although Vision-Language Models (VLMs) have shown significant advances in various multimodal tasks, there is still limited evidence regarding their ability to replicate specialized judgments in highly complex domains. In this work, we investigate the feasibility of using VLMs to annotate medicalization in advertisements from the Meta Ad Library. From a dataset of 7,167 ads containing text and images, we selected 300 examples that require specialized knowledge to assess their level of medicalization. The advertisements were annotated by three human experts and by three VLMs from different families, allowing for the comparison of human agreement, inter-model agreement, and the alignment between these groups. The results show that VLMs achieve agreement levels close to those observed among human experts in identifying medicalized content, suggesting that they are already capable of capturing a substantial portion of the criteria used by experts in this task. However, important differences are observed regarding the intensity attributed to the most ambiguous cases. Additionally, we identified substantial variability among models in aspects such as consistency across repeated runs and adherence to the annotation instructions. Finally, we conducted a qualitative analysis of the cases with the highest agreement and disagreement, characterizing the types of ads that facilitate and hinder automatic annotation. The results suggest that VLMs can serve as support for building specialized datasets for monitoring medicalization elements in advertisements, while also highlighting important limitations that must be considered in such applications.

KEYWORDS

Inteligência Artificial, *Visual-Language-models*, VLM, anúncios, *Ad Library*, dados de saúde, medicalização, Meta

1 INTRODUÇÃO

As plataformas digitais transformaram profundamente a circulação de informações sobre saúde [10], impulsionando a disseminação de

anúncios direcionados que frequentemente misturam conceitos médicos, de bem-estar, estética e intervenções clínicas. Nesse cenário, ganha destaque o fenômeno da medicalização: o enquadramento de condições cotidianas ou alterações estéticas como problemas médicos que exigem tratamento específico [6]. Embora anúncios médicos passem por uma regulamentação rigorosa com diretrizes específicas sobre o tipo de conteúdo e a forma como tais propagandas podem ser divulgadas para o público consumidor, tanto no Brasil [2, 7] como no resto do mundo [18], identificar possíveis violações nesse ecossistema e monitorar a proliferação e a intensidade desse tipo de publicidade em plataformas como as da Meta (controladora do Facebook, Instagram e WhatsApp) representa um desafio premente para a saúde pública contemporânea [41, 42]. Essa urgência se amplifica pela escala massiva de circulação de anúncios e pela complexidade inerente à interpretação conjunta de textos e imagens (multimodalidade) produzidos com fins persuasivos.

Tradicionalmente, a análise desse tipo de conteúdo exige a atuação de especialistas humanos (como pesquisadores e profissionais médicos) capazes de discernir nuances e aplicar critérios clínicos precisos [44]. Contudo, os altos custos operacionais e a limitada disponibilidade desses profissionais inviabilizam a construção de conjuntos de dados rotulados na escala necessária para o monitoramento contínuo da Web [45]. Diante dessa limitação, os recentes avanços em *Large Language Models* (LLMs) e, mais especificamente, em Modelos de Visão-Linguagem (VLMs), surgem como uma alternativa promissora para automatizar em escala a anotação de dados complexos. Entretanto, a utilização desses modelos como anotadores especializados ainda depende de evidências empíricas obtidas em cenários reais, nos quais conhecimento de domínio, subjetividade e multimodalidade coexistem. Além disso, a qualidade dos dados de treinamento e avaliação tem sido reconhecida como um dos principais fatores limitantes do desempenho desses modelos [23, 24, 29].

Embora o uso de inteligência artificial na área da Saúde venha ganhando cada vez mais destaque [30, 31], evidenciado por *benchmarks* como o *HealthBench*¹, focado na avaliação de habilidades médicas em interações baseadas em texto, a literatura aponta que a avaliação desses sistemas ainda depende fortemente de dados sintéticos e cenários idealizados [20]. Iniciativas recentes, como o *HealthBench Professional*, buscam aproximar a avaliação de situações clínicas reais, mas permanecem limitadas ao meio textual

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2026). Lavras/MG, Brazil. Porto Alegre: Brazilian Computer Society, 2026.

© 2026 SBC – Brazilian Computing Society.

ISSN 2966-2753

¹<https://openai.com/index/healthbench/>

[21]. Consequentemente, ainda são escassas as evidências sobre a capacidade desses modelos de reproduzir julgamentos especializados em tarefas multimodais que exigem interpretação conjunta de imagens e textos, conhecimento contextual e critérios subjetivos, especialmente em cenários orgânicos caracterizados pela elevada variabilidade do mundo real.

O presente trabalho busca preencher essa lacuna ao investigar até que ponto VLMs são capazes de reproduzir julgamentos especializados em um contexto orgânico e intrinsecamente multimodal. A medicalização em publicidade digital constitui um caso particularmente desafiador desse tipo de tarefa, pois exige a integração de evidências textuais e visuais com conhecimento médico para distinguir diferentes níveis de medicalização. Assim, mais do que avaliar o desempenho de VLMs em um problema específico, este estudo caracteriza seu potencial e suas limitações como anotadores especializados em um domínio altamente subjetivo. Para isso, são propostas as seguintes Questões de Pesquisa (QPs): **QP1:** *O quão consistentes são os VLMs na aderência aos requerimentos de tarefas multimodais complexas?* **QP2:** *O quão efetiva é a rotulação por VLMs em comparação com a concordância entre especialistas médicos?* **QP3:** *Como se caracterizam os anúncios que levam a maiores consensos e discordâncias entre os especialistas e os VLMs?*

Para investigar essas questões, foi construído um *dataset* multimodal composto por pares de imagem e texto extraídos da *Meta Ad Library*. A partir de um universo de 7.167 anúncios relacionados à saúde, foi selecionada uma amostra de 300 anúncios para comparação das anotações produzidas por três especialistas médicos e três VLMs de famílias distintas, com todos seguindo a mesma taxonomia ordinal de Medicalização (graus de 0 a 3). Para garantir um *ground truth* confiável, os especialistas passaram por um processo estruturado de treinamento e calibração, aumentando progressivamente sua concordância até atingir um nível estatisticamente satisfatório.

Os resultados mostram que, embora os VLMs alcancem elevada concordância com especialistas na identificação binária da presença de medicalização ($AC1 > 0,8$), seu desempenho diminui substancialmente quando a tarefa exige distinguir os diferentes níveis da taxonomia. Esses resultados indicam que os modelos já conseguem reproduzir julgamentos especializados em tarefas de triagem, mas ainda apresentam limitações importantes quando a decisão depende da interpretação da intensidade da medicalização. Através do *dataset* anotado e validado, este trabalho oferece uma caracterização empírica dos limites atuais dos VLMs como anotadores especializados em publicidade multimodal relacionada à saúde, fornecendo evidências para o desenvolvimento de futuras abordagens de anotação assistida por inteligência artificial.

2 TRABALHOS RELACIONADOS

A aplicação de inteligência artificial na área de saúde tem crescido de maneira notável nos últimos anos, em parte pelo grande volume de dados clínicos disponibilizados e também pelos avanços tecnológicos que podem ser observados em aprendizado profundo e modelos fundacionais. Inicialmente focados mais em classificação de imagens médicas, apoio ao diagnóstico e previsão de riscos clínicos, as soluções baseadas em inteligência artificial agora englobam atividades mais abrangentes, com raciocínio, interpretação de linguagem natural e integração de múltiplas fontes de informação.

Os *Large Language Models* (LLMs) surgiram como uma tecnologia promissora para aplicações médicas, modelos como Med-PaLM demonstraram desempenho competitivo em avaliações de conhecimento médico, aproximando do nível de especialistas humanos em determinadas tarefas de resposta a perguntas clínicas como mostrado no estudo [33]. O crescente interesse no tema levou a comunidade científica e a indústria a criar benchmarks especializados, como o HealthBench apresentado no artigo [3], criado para avaliar sistematicamente as capacidades de modelos de linguagem em cenários relacionados à saúde. Esses benchmarks buscam medir não apenas a precisão das respostas, mas também aspectos críticos como segurança e confiabilidade.

No entanto, estudos apontam limitações importantes para a adoção desses sistemas em ambientes reais, como o [20], argumenta que resultados obtidos em *benchmarks* podem não refletir adequadamente o desempenho em situações clínicas reais devido à dependência de dados sintéticos e à simplificação dos contextos avaliados. Iniciativas mais recentes têm buscado incorporar dados provenientes de interações reais entre profissionais de saúde e sistemas de IA, aproximando os processos de avaliação dos desafios encontrados na prática clínica.

A natureza multimodal da medicina tem impulsionado o desenvolvimento de Vision Language Models (VLMs), capazes de mesclar informações visuais e textuais em um mesmo raciocínio. Trabalhos como [22], [17] e [26, 43], demonstram o potencial desses modelos para analisar imagens médicas, relatórios clínicos e outros tipos de conteúdo multimodal.

Redes sociais, plataformas de publicidade e ecossistemas online produzem diariamente grandes volumes de conteúdo multimodal cuja interpretação frequentemente exige conhecimento especializado. Embora a maior parte da literatura tenha se concentrado em contextos clínicos tradicionais, como radiologia, patologia e apoio ao diagnóstico, existe um interesse crescente na utilização desses modelos para tarefas de análise de conteúdo relacionado à saúde em ambientes digitais. Sendo assim, compreender a capacidade dos VLMs de reproduzir julgamentos especializados torna-se uma questão relevante em vários âmbitos da medicina.

No entanto, ainda não existem estudos que investiguem a efetividade de VLMs na anotação de anúncios relacionados à saúde, esse trabalho analisa sua concordância com especialistas humanos na identificação de diferentes níveis de medicalização em conteúdos publicitários.

2.1 Anotação de Dados e Benchmarking de LMs

Hoje, uma das principais estratégias para avaliação do desempenho (*benchmarking*) de Modelos de Linguagem, multimodais ou não, se dá através de testes análogos a uma tarefa de anotação de dados. Testes desse tipo exigem a aderência a um conjunto de instruções e fornecem determinado material a ser classificado seguindo essa rubrica, verificando o total de respostas corretas em relação a um *ground-truth*.

Esse é o caso em [40] onde os modelos são avaliados através de milhares de questões multimodais distribuídas entre diferentes áreas do conhecimento, exigindo a interpretação conjunta de imagens e textos para a produção de uma resposta correta. De maneira

semelhante, benchmarks médicos recentes como o [3] e o [15] utilizam conjuntos de casos desenvolvidos por especialistas para avaliar a capacidade dos modelos de reproduzir julgamentos considerados adequados em contextos relacionados à saúde. Embora esses benchmarks sejam normalmente apresentados como instrumentos de avaliação de desempenho, sua estrutura é fundamentalmente equivalente a uma tarefa de anotação especializada, na qual o modelo recebe uma entrada, aplica um conjunto de critérios previamente definidos e produz uma classificação ou julgamento que pode ser comparado a uma referência humana.

O reconhecimento dessa semelhança motivou o surgimento de uma linha de pesquisa dedicada ao uso de Modelos de Linguagem como substitutos ou auxiliares de anotadores humanos. Trabalhos recentes demonstram que LLMs são capazes de reproduzir com elevados níveis de concordância tarefas de classificação textual, análise de sentimentos, identificação de desinformação e outras atividades tradicionalmente realizadas por especialistas ou trabalhadores de plataformas de anotação [13, 36]. Paralelamente, consolidou-se o paradigma conhecido como *LLM as a Judge*, no qual modelos são empregados para emitir avaliações sobre conteúdos produzidos por humanos ou por outros sistemas de IA, se aproximando cada vez mais do papel desempenhado por avaliadores especializados [46].

Apesar dos resultados promissores, a utilização desses modelos como anotadores levanta questões importantes relacionadas à consistência, robustez e confiabilidade de suas decisões. Diferentes estudos apontam que pequenas alterações de prompt, variações entre modelos ou mesmo execuções repetidas de uma mesma tarefa podem resultar em julgamentos distintos [4]. Essas limitações tornam-se particularmente relevantes em domínios especializados, nos quais os critérios de classificação dependem de conhecimento contextual complexo e apresentam níveis inerentes de subjetividade. Nesse cenário, avaliar a concordância entre modelos e especialistas humanos torna-se um passo fundamental para determinar a viabilidade do uso de LLMs e VLMs em tarefas de anotação especializada.

Embora a literatura sobre *LLM as a Judge* tenha crescido rapidamente, a maior parte dos estudos concentra-se em tarefas predominantemente textuais. Ainda são relativamente escassas as investigações sobre o uso de VLMs como anotadores especializados em cenários genuinamente multimodais, nos quais a interpretação correta depende da análise conjunta de imagens e textos. Essa lacuna é particularmente relevante em ambientes digitais contemporâneos, onde grande parte do conteúdo circula em formatos multimodais e exige conhecimento contextual para sua interpretação.

Diante desse cenário, foi investigado o potencial de diferentes VLMs para atuar como anotadores especializados de anúncios relacionados à saúde. Mais especificamente, suas capacidades de reproduzir os julgamentos de especialistas humanos em uma taxonomia de medicalização, avaliando aspectos como consistência, concordância intermodelo e alinhamento com especialistas. Buscando contribuir para a compreensão dos limites e oportunidades do uso de modelos multimodais na construção de datasets especializados e em tarefas de monitoramento automatizado de conteúdo.

2.2 O Uso de VLMs para Tarefas Especializadas

Recentes avanços em inteligência artificial são impulsionados pelo crescimento de modelos multimodais, capazes de processar e integrar informações de diferentes modalidades de dados. Entre eles, os VLMs se destacam por combinar representações visuais e textuais em um espaço semântico compartilhado, possibilitando a realização de tarefas como descrição automática de imagens, resposta a perguntas visuais e recuperação multimodal de informações.

Os primeiros avanços foram inicialmente impulsionados por modelos como [27], alinhando representações de imagens e textos por meio de aprendizado contrastivo, utilizando grandes quantidades de dados disponíveis na Web. Arquiteturas como [1] e [19] ampliaram as capacidades desses sistemas, possibilitando interações mais próximas das observadas em modelos conversacionais e incorporando mecanismos de instrução em linguagem natural.

Em áreas como saúde, a correta interpretação dos dados vai geralmente além da identificação de padrões visuais ou textuais, exigindo também conhecimentos específicos e frequentemente implícitos, alheios a leigos no assunto. Neste sentido, a literatura ainda apresenta poucas evidências sobre a capacidade desses modelos de reproduzir julgamentos especializados em tarefas que envolvem elementos subjetivos, normativos ou socioculturais. O presente trabalho contribui para a literatura ao avaliar o desempenho de diferentes VLMs em uma tarefa especializada de anotação de anúncios relacionados à saúde, um domínio que combina conhecimento médico, interpretação contextual e análise conjunta de imagens e textos. Ao comparar os resultados produzidos pelos modelos com os de especialistas humanos, busca-se compreender o potencial e as limitações desses sistemas para apoiar a construção de datasets especializados e processos de monitoramento em larga escala.

3 METODOLOGIA

Esta seção detalha o fluxo de trabalho adotado na construção e análise do dataset especializado em anúncios médicos. A metodologia conduzida neste estudo visa contrastar sistematicamente a capacidade de anotação de especialistas humanos com a de modelos computacionais em um domínio restrito e altamente subjetivo. Para tanto, estruturou-se uma abordagem em quatro etapas fundamentais, conforme ilustrado no fluxograma da Figura 1: coleta na Ad Library, construção do dataset (incluindo pré-filtragem), rotulação por especialistas humanos e avaliação via modelos de Visão-Linguagem (VLMs) para fins de comparação empírica. A seguir, detalhamos cada uma dessas fases.

3.1 Coleta de Conteúdo na Meta Ad Library

Como fonte de dados, foi utilizada a plataforma de anúncios Meta Ad Library: um repositório oficial concebido como uma iniciativa de transparência, que viabiliza o acesso público aos anúncios veiculados nas diversas plataformas da empresa (como Facebook e Instagram) [11]. Embora seja uma ferramenta de grande importância para compreender o ecossistema de anúncios, o acesso aos dados é condicionado por legislações locais e políticas de transparência da plataforma, que frequentemente apresentam inconsistências [32], esbarram na falta de transparência da infraestrutura e em restrições de acesso a dados completos [16]. No Brasil, por exemplo, o detalhamento adicional de gastos, impressões e audiência é restrito

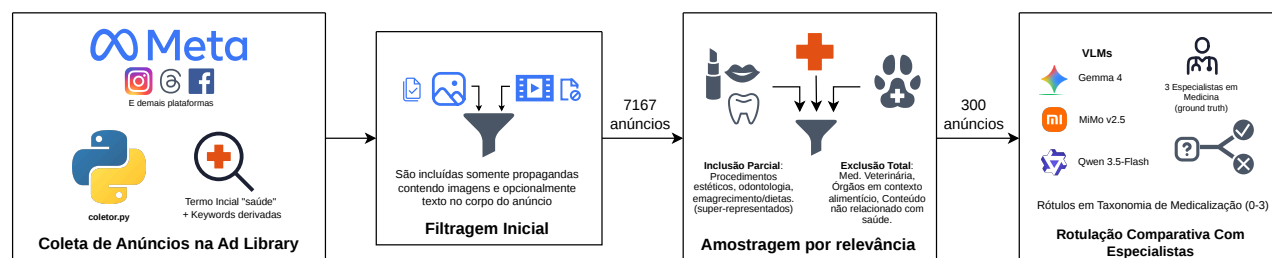


Figura 1: Fluxograma da metodologia geral de coleta, anotação e análise dos anúncios médicos.

a anúncios classificados como “políticos”, havendo menor visibilidade em categorias comerciais, incluindo a área da saúde. Para os dados deste estudo, o maior impacto dessa limitação técnica foi a impossibilidade de se extrair as quantidades precisas do total de anúncios em circulação associados às palavras-chave (*keywords*) utilizadas. A busca pela palavra-chave semente “saúde”, por exemplo, sempre retornava o teto limitador da API (50.000 resultados), indicando apenas que o volume real era superior a essa marca, sem revelar o quantitativo exato. Diante dessa limitação, a estratégia de extração via documentação da API² precisou ser adaptada para uma abordagem iterativa e amostral.

Para construir um *dataset* representativo, foi definido um conjunto de 33 palavras-chave derivadas iterativamente da palavra semente “saúde”. A identificação de elementos informativos em conteúdo Web tem sido explorada em recuperação de informação, incluindo abordagens baseadas na estimativa da importância de blocos de conteúdo [12]. Devido à alta incidência de anúncios ruidosos ou irrelevantes no retorno inicial, novas *keywords* foram sendo selecionadas com base nos termos médicos que coocorriam com maior frequência nos artefatos relevantes. Esse processo de coleta orgânica resultou em uma base inicial de 7.167 anúncios.

Com o objetivo de otimizar o esforço do comitê de especialistas, o *dataset* foi submetido a uma etapa de pré-filtragem para remover conteúdos evidentemente fora do escopo da saúde humana, mas que compartilhavam léxico semelhante. Estratégias de seleção de instâncias têm sido exploradas para reduzir redundância e ruído nos dados, preservando a efetividade e reduzindo custos de processamento [8, 9]. Foram excluídos, por exemplo, anúncios de serviços veterinários, ofertas de supermercados (que coincidiam com termos anatômicos como “fígado” para a venda de carnes) e usos irrelevantes de termos médicos (ex: “vírus” de computador ou “dor” em narrativas de ficção).

Dessa base filtrada, selecionou-se uma amostra final de 300 anúncios para o processo de anotação mais granular. A modelagem de tópicos tem explorado diferentes formas de incorporar relações semânticas às representações textuais, incluindo representações baseadas em *word embeddings*, estruturas hierárquicas e *contextual embeddings* [37–39]. A coesão temática e os tópicos prevalentes no *dataset* (tanto os excluídos pela pré-filtragem quanto os incluídos na amostragem final) foram mapeados utilizando o algoritmo de

Tabela 1: Tópicos gerados via aplicação do BERTopic a partir de cada amostra de anúncios coletados.

Tópico	Excluídos	Incluídos: 300
0	sorriso, implante, dentes tratamento, prótese, estética agende, avaliação, implantes, sorrir	dia, saúde, corpo, pode resultados, gordura, agora detox, metabolismo, forma
1	pet, saúde, animal, felinos vacina, atendimento, veterinário pode, vet, whatsapp	dor, dores, terapia ansiedade, vida, tratamento corpo, pode, ajuda, saúde
2	ofertas, supermercado, sábado qualidade, melhor, ofertas válidas whatsapp, válidas, preços, economia	vacina, câncer, saúde, hpv exame, prevenção, cuidado vacinas, meses, proteção
3	creatina, whey, saúde, muscular alimentação, strong, sódio strong nutrition, nutrition whey protein	saúde, curso, atendimento homeopatia, whatsapp, clínica profissional, medicina, clique botão
4	emocional, depressão, ansiedade terapia, experiências, apego dependência, relações, vida, mental	pele, agende, tratamento saúde, avaliação, autoestima sorriso, dr, dentes, pode

modelagem de tópicos BERTopic, cujos agrupamentos estão detalhados na Tabela 1. A diferença temática evidencia a eficácia da pré-filtragem em remover dados ruidosos (representados à esquerda) para garantir a coesão clínica da amostra final (à direita).

3.2 Rotulação Humana Médica Especializada

Para suprir esses requisitos operacionais foi utilizada a *Labelbox*, uma plataforma amplamente utilizado para tal tarefa [25]. A literatura aponta que a anotação manual de dados de saúde exige considerável rigor bioético, além de conhecimento de termos técnicos e normas [45]. Adicionando a isso, o contexto multimodal apresenta uma complexidade cognitiva e técnica substancialmente maior do que tarefas puramente textuais, exigindo ambientes de rotulação otimizados e bem definidos [28, 35].

Para a condução desta etapa, o painel de anotação foi composto por três pesquisadores médicos especialistas, previamente familiarizados com o escopo e o contexto do estudo. A tarefa consistiu na avaliação minuciosa dos anúncios (pares de imagem e texto) para classificar o nível de medicalização do conteúdo veiculado. Os especialistas foram instruídos a atribuir um grau de classificação e fornecer uma justificativa textual baseada na interpretação do anúncio, sendo expressamente orientados a aplicar as definições da taxonomia ao contexto visualizado. A tarefa exata (tanto para os especialistas médicos quanto às VLM) é apresentada no prompt abaixo. A avaliação seguiu uma taxonomia de medicalização estruturada em quatro níveis incrementais:

²<https://www.facebook.com/ads/library/api>

- **Grau 0 (Não medicalizado):** Conteúdos que não apresentam enquadramento médico (sem referência a doenças, sintomas ou condições clínicas), com linguagem restrita a bem-estar, estética ou qualidade de vida.
- **Grau 1 (Leve):** Uso de linguagem genérica relacionada à saúde, mas de forma ampla e inespecífica, sem indicar implícita ou explicitamente uma doença, condição específica ou disfunção.
- **Grau 2 (Moderado):** Conteúdos que introduzem a possibilidade de alteração, disfunção ou risco fisiológico. A anormalidade é sugerida, ainda que de forma indireta, sem afirmar categoricamente a presença de uma doença.
- **Grau 3 (Forte):** Enquadramento explícito de uma condição como problema médico ou doença. A linguagem caracteriza a situação clinicamente, frequentemente associada à sugestão de necessidade de intervenção.

3.3 Rotulação via Modelos de Visão-Linguagem (VLMs)

Para testar a viabilidade da automação desta tarefa, uma mesma amostra com exatamente a mesma tarefa foi submetida a três modelos multimodais recentes. A inferência foi orquestrada por meio da API unificada *OpenRouter*³, que permite o roteamento simultâneo de um mesmo *prompt* para múltiplos modelos, garantindo a recepção de instruções estritamente idênticas e facilitando o gerenciamento dos dados. A escolha dos modelos baseou-se na sua escalabilidade técnica e financeira para o escopo deste estudo. Dessa forma, selecionamos VLMs de pesos abertos (*open-weights*) que oferecem o menor custo de *input/output* por milhão de *tokens*, mas que reportam desempenho competitivo em *benchmarks* multimodais complexos, como o avaliado no estudo MMMU [40]. Os modelos inferidos foram o Qwen3.5-Flash, MiMo v2.5 e Gemma 4.

3.4 Engenharia de Prompt e Taxonomia de Medicalização

A estrutura do prompt foi construída buscando minimizar diferenças em relação às instruções fornecidas por escrito na plataforma de rotulação. A única adaptação realizada foi a introdução de delimitadores de formatação necessários para tarefas de rotulação padronizada via IAs generativas, destacados em negrito no prompt abaixo.

Classifique o anúncio exibido de acordo com os critérios de Medicalização definidos abaixo, seguindo essa mesma ordem e utilizando em cada critério o seguinte formato, sem incluir markdown:

Veredito [Critério]: Grau 0-3 Justificativa: [Justificativa]

A resposta deve se basear nos critérios das respectivas dimensões, podendo mencioná-los de acordo com o contexto, mas sendo proibido repetir a totalidade de qualquer uma das definições.

Critérios Medicalização:

Grau 0 – Não medicalizado Definição: Inclui conteúdos que não apresentam enquadramento médico, ou seja, não há referência a doença, sintoma ou condição clínica. A linguagem é restrita a aspectos gerais de bem-estar, qualidade de vida ou estética. Atenção: Não pode incluir qualquer sugestão, explícita ou implícita, de problema fisiológico ou alteração de saúde. Exemplo: "Produto natural para melhorar sua disposição."

Grau 1 – Leve Definição: Inclui conteúdos que utilizam linguagem genérica relacionada à saúde, sem implicar a existência de uma condição específica ou disfunção. A referência à saúde é ampla e inespecífica. Atenção: Não há indicação implícita de doença ou alteração patológica. Exemplo: "Ajuda na saúde do fígado."

Grau 2 – Moderado Definição: Inclui conteúdos que sugerem a possibilidade de alteração, disfunção ou risco, sem afirmar diretamente a presença de doença. Há introdução de anormalidade, ainda que de forma indireta. Atenção: Introdução de risco, disfunção ou anormalidade fisiológica. Exemplo: "Cansaço pode indicar problemas no fígado."

Grau 3 – Forte Definição: Inclui conteúdos que apresentam enquadramento explícito de uma condição como problema médico, associado à sugestão de necessidade de intervenção. A linguagem já caracteriza a situação como doença ou condição clínica. Atenção: Definição explícita do problema como condição médica. Exemplo: "Trate sua doença hepática."

3.5 Análise de Dados

A extração de significado estatístico e semântico dos dados apoiou-se em métricas consolidadas na literatura para garantir reprodutibilidade. **Métricas de Concordância:** Para calcular as concordâncias internas e externas entre os grupos humanos e algorítmicos, adotamos a métrica Gwet's AC1 (Autoconcordância de Primeira Ordem). Esta métrica foi selecionada por sua robustez e resistência a paradoxos estatísticos comuns ao Kappa de Cohen, sendo especialmente adequada para casos de alto desbalanceamento no total das categorias, como o observado no espectro de medicalização do nosso *dataset*.

Modelagem de Tópicos: A extração não-supervisionada dos temas presentes nas campanhas foi realizada por meio do *framework* BERTopic [14]. Tratando-se de uma ferramenta modular que utiliza um mecanismo de atenção baseado em *Transformers*, optamos por incorporar o *embedding paraphrase-multilingual-MiniLM-L12-v2*. Essa escolha técnica justifica-se pela sua alta otimização e acurácia na vetorização de textos curtos e incisivos, cenário característico das legendas publicitárias em redes sociais.

4 RESULTADOS E DISCUSSÕES

Nesta seção, apresentamos os resultados da avaliação comparativa entre especialistas humanos e Modelos de Visão-Linguagem (VLMs). A análise está estruturada em torno das três questões de pesquisa (QPs) norteadoras deste estudo, abordando primeiramente a consistência interna dos modelos, seguida pela avaliação de sua efetividade em relação ao *ground-truth* humano e, por fim, uma análise qualitativa dos cenários de consenso e discordância.

4.1 Consistência e Aderência dos VLMs (QP1)

Uma das principais barreiras para a adoção de Modelos de Visão-Linguagem em tarefas de auditoria regulatória e anotação sistemática de dados é o não-determinismo inerente às arquiteturas generativas modernas, persistente mesmo sob configurações de temperatura nula ($T = 0$) [5]. No contexto da construção de *datasets* especializados, a variabilidade das respostas pode comprometer a validade científica das análises subsequentes. Dessa forma, seguindo as recomendações metodológicas de [34], a primeira dimensão avaliada neste estudo foi a consistência de cada modelo ao realizar três execuções redundantes para cada anúncio da amostra. Os resultados de autoconcordância estão sintetizados na Tabela 2.

³<https://openrouter.ai/>

A análise da Tabela 2 expõe diferenças substanciais entre as famílias de modelos. O modelo *Gemma 4 (26b)* apresentou um desempenho técnico superior nas duas dimensões avaliadas, com um escore AC1 de autoconcordância interna de 0.83034 e falhou na estruturação de saída em apenas uma das 300 tarefas analisadas (299/300 de sucesso de formatação). Esse resultado ainda foi alcançado com a menor pegada de custo relativo (12.7%), consolidando sua viabilidade operacional para tarefas em larga escala.

Em contrapartida, o modelo *MiMo v2.5* revelou-se o mais instável, com um AC1 interno de apenas 0.49576 e desvio de restrições de formatação em 80 tarefas (220/300 de sucesso), levando ao custo relativo mais alto de 55.5%. O *Qwen3.5-Flash* ocupou um espaço intermediário, com autoconcordância moderada de 0.66110 e conformidade de formatação de 269/300. Essa instabilidade do *Qwen3.5-Flash* é esperada de variantes otimizadas para baixa latência (modelos “Flash”).

Para avaliar a precisão dos modelos quanto à nossa taxonomia, foi realizado um experimento de ablação comparando um prompt de controle (instruções básicas) com um prompt completo (contendo as definições detalhadas de cada grau de medicalização). Os resultados, apresentados na Tabela 3, demonstram que a inclusão dessas diretrizes detalhadas dobrou tanto a concordância entre os modelos quanto o alinhamento destes com os especialistas humanos. Isso comprova que, para classificar fenômenos subjetivos, os VLMs atuais dependem de instruções explícitas, não alcançando o desempenho desejado de forma espontânea (zero-shot).

4.2 Eficácia da Rotulação e Alinhamento com Especialistas (QP2)

Para responder à QP2, avaliou-se o nível de concordância externa entre as anotações geradas pelos VLMs e o padrão-ouro estabelecido pela rotulação dos médicos especialistas. A Tabela 4 apresenta uma comparação detalhada dos escores AC1 de Gwet calculados tanto na escala ordinal completa (Graus 0 a 3) quanto na escala binarizada (presença versus ausência de medicalização).

Os dados da Tabela 4 revelam um comportamento bem distinto na eficácia dos modelos que depende da resolução da classificação. Quando a tarefa é binarizada (agrupando Grau 0 como “Não medicalizado” e os Graus 1, 2 e 3 como “Medicalizado”), os VLMs alcançam níveis de alinhamento altos com os especialistas humanos, com todos os valores superando a barreira de 0.78. Ainda, a incorporação de VLMs ao conjunto de anotações binarizadas dos especialistas humanos aumentou o escore AC1 global do grupo para 0.81821 (em relação ao escore inter-humanos de 0.78778). Esse incremento estatístico sugere que a votação majoritária assistida por IA atuou como um mecanismo de desempate eficaz, estabilizando, de certa forma, as divergências observadas entre os anotadores humanos. Ao analisarmos o desempenho individual de alinhamento com os médicos humanos, o *Qwen3.5-Flash* obteve a maior concordância (0.81112), superando o *Gemma 4* (0.79931) e o *MiMo v2.5* (0.78389). Esse resultado sugere uma revelação metodológica contraintuitiva,

Tabela 2: Comparação da consistência interna dos modelos.

Métrica de Consistência	Qwen3.5-Flash	MiMo v2.5	Gemma 4 (26b)
AC1 Score (Autoconcordância)	0.66110	0.49576	0.83034
Total de Sucesso (Aderência Formato)	269/300	220/300	299/300
Pegada de Custo Relativo (%)	31.7%	55.5%	12.7%

Tabela 3: Efeito da estruturação especializada do *prompt* na concordância inter-modelos, calculado sobre uma subamostra de 60 tarefas.

Conjunto de Rotuladores	Prompt	Escore AC1
Modelos de Visão-Linguagem Especialistas Humanos + VLMs	Controle	0.25579 0.18369
Modelos de Visão-Linguagem Especialistas Humanos + VLMs	Completo	0.51344 0.50533

onde a capacidade de um modelo de manter a consistência de formatação sintática e fidelidade interna inter-rodadas (Tabela 2) não se traduziu na maior acurácia.

Entretanto, quando a tarefa exige a granulação fina da escala completa (diferenciando os níveis de intensidade da medicalização), observa-se um declínio acentuado nos escores de concordância de todos os grupos. Mesmo a concordância inter-especialistas humanos cai para 0.50533. O alinhamento dos modelos individuais com os especialistas na escala ordinal seguiu essa queda, com valores entre 0.44 a 0.47. A distribuição de escolhas dos modelos para os 68 casos de consenso absoluto dos especialistas (CE) está graficamente ilustrada na Figura 3. O gráfico revela que os modelos demonstram proficiência em mapear e replicar as decisões humanas nos extremos da taxonomia (especialmente no Grau 0), mas sofrem de um severo ruído de calibração nas categorias intermediárias, tendendo a classificar de forma desbalanceada e pulverizada os anúncios de Grau 1 e Grau 2.

4.3 Análise Qualitativa de Consensos e Discordâncias Multimodais (QP3)

A caracterização temática dos anúncios e seu respectivo alinhamento foi investigada qualitativamente a partir dos agrupamentos temáticos do BERTopic expostos na Tabela 5 e nos anúncios com maior discordância exibidos na Figura 2.

A caracterização dos dados evidencia que as tarefas que facilitam a automação da anotação das VLMs aquelas em que os eixos visual e textual estão mais alinhados e apresentam sinais médicos mais evidentes. Conforme observado na Tabela 5, o Tópico 9 (caracterizado por termos de patologias como *câncer*, *HPV*, *vírus* e *exame*) e o

Tabela 4: Comparação dos escores de concordância para diferentes combinações de rotuladores humanos e artificiais.

Conjunto de Rotuladores	Escore Gwet's AC1	
	Escala Ordinal (0–3)	Escala Binarizada
Especialistas Humanos	0.50533	0.78778
Modelos de Visão-Linguagem (Consenso)	0.54482	0.87512
Especialistas + VLMs (Grupo Geral)	0.45300	0.81821
Especialistas + Qwen3.5-Flash	0.44476	0.81112
Especialistas + MiMo v2.5	0.44598	0.78389
Especialistas + Gemma 4 (26b)	0.47461	0.79931

Tabela 5: Modelagem de tópicos dos anúncios onde ocorreu alinhamento entre VLMs e o consenso especialista.

Tópico	Termos e Vocabulário Prevalente (BERTopic)
0	dia saúde resultados pode gordura corpo emagrecimento agora metabolismo acompanhamento
8	whatsapp clinica doutor doutor saúde exame clínica consulta exame botão envie mensagem whatsapp
3	dia saúde resultados pode gordura corpo emagrecimento
9	câncer hpv exame exames saúde diagnóstico cada principais homens vírus
-1	terapia ansiedade depressão emocional autoconhecimento vida ansiedade depressão

Tópicos da Discordância (BERTopic)

-1	natural remédio natural remédio saúde pode dia corpo anos agora vida
2	pele fios tecnologia cabelo tratamento cuidar resultados agende procedimento lase
0	dia saúde resultados pode gordura corpo emagrecimento agora metabolismo acompanhamento
9	câncer hpv exame exames saúde diagnóstico cada principais homens vírus
4	terapia ansiedade depressão emocional autoconhecimento vida ansiedade depressão saúde mental holística



Figura 2: Anúncios onde ocorreram os maiores distanciamentos entre VLMs e o consenso humano, com as categorias utilizando o mesmo esquema de cores das categorias de medicalização da Figura 3. A presença de uma célula amarela em meio a verdes indica uma discordância binária. Respectivamente: 1. O único caso de opinião absolutamente oposta, com unanimidade 0 entre os especialistas e consenso 3 entre os modelos.

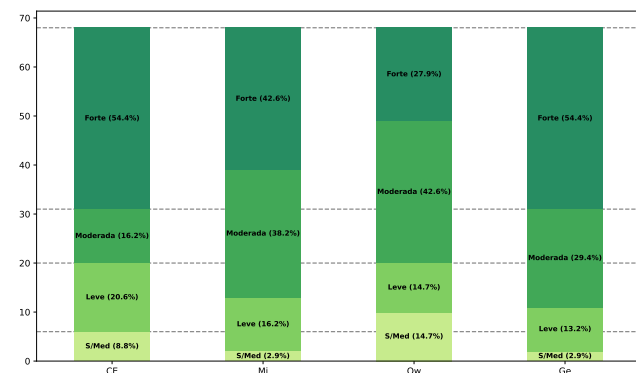


Figura 3: Comparação da distribuição de escolhas de categoria nos casos em que houve consenso absoluto entre especialistas (CE), totalizando 68 casos. As linhas pontilhadas indicam a proporção das categorias em CE.

Tópico 8 (consultas, agendamentos e clínicas médicas) obtiveram alto consenso.

No lado oposto, correspondente à não-medicalização absoluta (Grau 0), o Tópico 0 e o Tópico 3 (focados em condicionamento físico, estética corporal e nutrição de bem-estar) também alcançam elevado consenso. Anúncios de suplementos ou programas de emagrecimento geral que não realizam alegações terapêuticas ou patologizações são confiantemente categorizados como Grau 0 por ambos os rotuladores, demonstrando que os VLMs capturam com precisão os limites entre promoção estética geral e intervenções de saúde.

O verdadeiro ponto de atrito entre VLMs e especialistas reside nas categorias em que a medicalização é implícita ou quando envolve tópicos de saúde mental (Tópico -1). De forma notável, o tópico -1 reúne os termos em que BERTopic também teve dificuldades no alinhamento. A apropriação comercial e a banalização contemporânea de termos como *terapia*, *ansiedade* e *depressão* em redes sociais criam dificuldades semânticas para as IAs.

A dissonância multimodal intencional representa o ponto de maior desafio lógico das VLMs, como visto nos exemplos da Figura 2: **O Caso da Notícia Hospitalar:** O único anúncio em que se observou discordância binária absoluta entre especialistas e VLMs corresponde a uma notificação de um acidente que resultou em uma

infecção hospitalar. Apesar dos termos médicos explícitos, o anúncio não os atribui ou direciona para a audiência, o que demonstra sua não-medicalização. Esse caso de falha, observado em todos os modelos, demonstra que nem sempre essas tecnologias são capazes de compreender a totalidade contextual no mesmo nível que humanos.

O Caso do composto capilar natural: No anúncio do “Kit Óleos”, a imagem apresenta elementos inofensivos, sugerindo um cosmético comum. Contudo, o texto menciona um “detox natural dos fios”, um termo cujas implicações fisiológicas são dependentes do contexto, algo que se reflete na total discordância classificatória entre os modelos. Como pode ser observado, apenas o modelo Qwen-3.5 chegou na mesma conclusão dos especialistas, onde se observou o que, apesar da ambiguidade, não há enquadramento da condição como doença. Este exemplo demonstra as dificuldades de convergência dos VLMs em contextos ambíguos.

O Caso Ozempic em Cápsulas: O anúncio explora a marca farmacêutica mais famosa da atualidade para vender um suplemento de emagrecimento duvidoso. O comitê de especialistas identificou a fraude regulatória de se anunciar um hormônio injetável de prescrição restrita como pílulas milagrosas de venda direta em redes sociais, mas rotulando-o como Grau 1 sobre medicalização. Em contrapartida, os VLMs entram em confusão: detectam a entidade nomeada de alto risco e classificam unanimemente a peça no Grau 3, focando na intenção fraudulenta e a apropriação predatória do anúncio. Outro exemplo dessa discordância de Grau 3 é o anúncio do **Cirurgião Bucomaxilofacial**. Multimodalmente, o anúncio não deixa muita margem para subjetividade: o texto estipula a realização de exames diagnósticos, cirurgias ortognáticas e tratamentos para tumores, enquanto a imagem exibe uma foto de alta credibilidade clínica com texto e planos de saúde. Diante dessa convergência os especialistas médicos classificam unanimemente o anúncio como Grau 3, enquanto algumas VLMs tiveram dificuldade de perceber tais elementos.

Essas descobertas qualitativas confirmam que as falhas estruturais das IAs não residem na interpretação literal das palavras-chave, mas no desconhecimento de dinâmicas mercadológicas, táticas de evasão semântica e na incapacidade de articular julgamento pragmático sob ambientes mutáveis de publicidade digital.

5 CONCLUSÃO

Este trabalho investigou a capacidade de Modelos de Visão-Linguagem (VLMs) de atuar como anotadores especializados em uma tarefa multimodal de elevada complexidade: a identificação de diferentes níveis de medicalização em anúncios relacionados à saúde. Diferentemente de avaliações conduzidas em *benchmarks* sintéticos ou predominantemente textuais, nosso estudo foi realizado em um contexto orgânico, composto por anúncios reais da *Meta Ad Library*, nos quais a decisão depende da integração entre linguagem, elementos visuais e conhecimento médico especializado.

Os resultados mostram que os VLMs já são capazes de reproduzir, com elevado grau de concordância, julgamentos especializados em tarefas de triagem, distinguindo de forma consistente a presença ou ausência de medicalização. Entretanto, quando a tarefa exige discriminar diferentes níveis de intensidade do fenômeno,

observa-se uma redução significativa da concordância com especialistas humanos. A análise qualitativa revelou que essa limitação decorre principalmente da dificuldade dos modelos em interpretar aspectos pragmáticos e normativos presentes na publicidade em saúde, especialmente em anúncios que combinam linguagem persuasiva, promessas de bem-estar e referências implícitas a condições médicas.

Esses resultados sugerem que a principal limitação atual dos VLMs não está na identificação de conceitos médicos explícitos, mas na reprodução de julgamentos especializados cuja decisão depende de inferências contextuais e critérios subjetivos. Assim, embora esses modelos já apresentem potencial para apoiar a construção de conjuntos de dados especializados e tarefas de triagem em larga escala, a supervisão humana permanece fundamental em cenários que exigem interpretações mais refinadas.

Além da caracterização empírica dessas limitações, este trabalho disponibiliza um *dataset* multimodal anotado e validado por especialistas médicos, contribuindo para futuras pesquisas sobre anotação especializada, avaliação de VLMs e construção de *benchmarks* mais representativos de cenários reais. Espera-se que os resultados apresentados incentivem o desenvolvimento de modelos multimodais capazes não apenas de reconhecer conceitos médicos, mas também de incorporar aspectos contextuais e subjetivos inerentes aos processos de decisão humana.

5.1 Limitações e Trabalhos Futuros

Este estudo apresenta algumas limitações. Primeiramente, a avaliação foi conduzida em uma amostra de 300 anúncios, suficiente para uma análise comparativa controlada, mas ainda limitada frente à diversidade de estratégias de publicidade encontradas em plataformas digitais. Além disso, foram avaliadas apenas três famílias de VLMs, todas selecionadas por apresentarem uma boa relação entre custo e desempenho para cenários de anotação em larga escala.

Como trabalhos futuros, pretende-se ampliar o conjunto de anúncios anotados, incorporar modelos multimodais mais recentes e investigar estratégias híbridas de anotação, nas quais especialistas humanos e VLMs atuem de forma complementar. Outra direção promissora consiste em explorar mecanismos de calibração e explicabilidade que permitam compreender melhor os fatores responsáveis pelas discordâncias observadas entre modelos e especialistas, contribuindo para o desenvolvimento de sistemas de anotação assistida mais confiáveis para aplicações em saúde.

CUIDADOS ÉTICOS

Em conformidade com a Lei Geral da Proteção de Dados (Lei nº 13.709/2018), o presente estudo não utiliza dados de privados em sua investigação, com a divulgação para o público sendo o propósito fundamental por trás de propagandas. A aderência aos termos de utilização dos serviços de anúncios Meta implica no registro dos anúncios na Biblioteca, onde ficam acessíveis ao público, também impondo a responsabilização quanto ao conteúdo incluído neles aos próprios anunciantes⁴. Ainda sim, com o intuito de promover boas práticas de ética em pesquisa, foi realizada a anonimização de nomes próprios e rostos incluídos nas propagandas, tanto para

⁴<https://pt-br.facebook.com/business/help/327111418314780/>

terceiros mencionados quanto para os próprios anunciantes, como visto na Figura 2.

Ferramentas de IA generativa foram aplicadas apenas na revisão gramatical e clareza do texto. A concepção do estudo, metodologia, experimentos, análise dos resultados e conclusões são de responsabilidade integral dos autores.

AGRADECIMENTOS

Este trabalho foi apoiado pelo CNPq, CAPES, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILDIAR, bolsa n° 408490/2024-1), Centro de Inovação e Inteligência Artificial para Saúde (CI-IA Saúde), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, Processo n° 2020/09866-4), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, Processo n° PPE-00030-21 e Projeto n° APQ-03115-24), UNIMED Belo Horizonte, AWS e NVIDIA.

REFERÊNCIAS

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* 35 (2022), 23716–23736.
- [2] ANVISA. 2008. RESOLUÇÃO-RDC N° 96, DE 17 DE DEZEMBRO DE 2008. Dispõe sobre a propaganda, publicidade, informação e outras práticas cujo objetivo seja a divulgação ou promoção comercial de medicamentos.
- [3] Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. HealthBench: Evaluating Large Language Models Towards Improved Human Health. *arXiv:2505.08775* [cs.CL]. <https://arxiv.org/abs/2505.08775>
- [4] Merve Astekin, Max Hort, and Leon Moonen. 2024. An exploratory study on how non-determinism in large language models affects log parsing. In *Proceedings of the ACM/IEEE 2nd International Workshop on Interpretability, Robustness, and Benchmarking in Neural Software Engineering*. 13–18.
- [5] Merve Astekin, Max Hort, and Leon Moonen. 2024. An Exploratory Study on How Non-Determinism in Large Language Models Affects Log Parsing. In *Proceedings of the ACM/IEEE 2nd International Workshop on Interpretability, Robustness, and Benchmarking in Neural Software Engineering* (Lisbon, Portugal) (InteNSE '24). Association for Computing Machinery, New York, NY, USA, 13–18. doi:10.1145/3643661.3643952
- [6] Peter Conrad. 2007. *The medicalization of society: On the transformation of human conditions into treatable disorders*. Vol. 14. Johns Hopkins University Press Baltimore.
- [7] Conselho Federal de Medicina. 2023. Resolução CFM n° 2.336/2023. Dispõe sobre publicidade e propaganda médicas. Diário Oficial da União, Brasília, DF, n° 175, 13 set. 2023. Disponível em: <https://www.in.gov.br/web/dou/-/resolucao-cfm-n-2.336-de-13-de-julho-de-2023-509398401>.
- [8] Washington Cunha, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2023. An Effective, Efficient, and Scalable Confidence-Based Instance Selection Framework for Transformer-Based Text Classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 665–674. doi:10.1145/3539618.3591638
- [9] Washington Cunha, Felipe Viegas, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2025. A Noise-Oriented and Redundancy-Aware Instance Selection Framework. *ACM Transactions on Information Systems* 43, 1 (2025). doi:10.1145/3701769
- [10] Jonas Henrique dos Santos, Julio CS Reis, Philippe de Freitas Melo, João Francisco Hecksher Olivetti, Thales Henrique Silva, Matheus Gontijo Guimaraes, Glauco de Souza, Marcos Gonçalves, Fabricio Benevenuto, Felipe Belchior Bessa Zanovello, et al. 2026. Whatsapp vaccine discourse (whavax): An expert-annotated dataset and benchmark for health misinformation detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 20. 2768–2779.
- [11] Laura Edelson, Tobias Lauinger, and Damon McCoy. 2020. A Security Analysis of the Facebook Ad Library. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 661–678. doi:10.1109/SP40000.2020.00084
- [12] David Fernandes, Edleno Silva de Moura, Berthier A. Ribeiro-Neto, Altigran Soares da Silva, and Marcos André Gonçalves. 2007. Computing Block Importance for Searching on Web Sites. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*. ACM, 165–174. doi:10.1145/1321440.1321466
- [13] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120, 30 (2023), e2305016120.
- [14] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).
- [15] Rebecca Soskin Hicks, Mikhail Trofimov, Dominick Lim, Rahul K Arora, Foivos Tsimpourlas, Preston Bowman, Michael Sharman, Chi Tong, Kavin Karthik, Arnav Dugar, et al. 2026. HealthBench Professional: Evaluating Large Language Models on Real Clinician Chats. *arXiv preprint arXiv:2604.27470* (2026).
- [16] Victor Le Pochat, Laura Edelson, Tom Van Goethem, Wouter Joosen, Damon McCoy, and Tobias Lauinger. 2022. An Audit of Facebook's Political Ad Policy Enforcement. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, 607–624.
- [17] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890* (2023).
- [18] Bryan A Liang and Tim Mackey. 2011. Direct-to-consumer advertising with interactive internet media: global regulation and public health issues. *JAMA* 305, 8 (2011), 824–825.
- [19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems* 36 (2023), 34892–34916.
- [20] Jialin Liu and Siru Liu. 2025. HealthBench: Advancing AI evaluation in healthcare, but not yet clinically ready. *DIGITAL HEALTH* 11 (2025), 20552076251390447. doi:10.1177/20552076251390447
- [21] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-Flamingo: a Multimodal Medical Few-shot Learner. In *Proceedings of the 3rd Machine Learning for Health Symposium (Proceedings of Machine Learning Research, Vol. 225)*, Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvinder Singh (Eds.). PMLR, 353–367.
- [22] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023. Med-Flamingo: A Multimodal Medical Few-shot Learner. (July 2023). <https://arxiv.org/abs/2307.15189> arXiv:2307.15189.
- [23] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Noumane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling Data-Constrained Language Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 50358–50376. https://proceedings.neurips.cc/paper_files/paper/2023/file/9d89448b63ce1e2e8dc7af72c984c196-Paper-Conference.pdf
- [24] Amanda Myntti, Erik Henriksson, Veronika Laipala, and Sampo Pyysalo. 2025. Register Always Matters: Analysis of LLM Pretraining Data Through the Lens of Language Variation. In *Second Conference on Language Modeling*. <https://openreview.net/forum?id=FqXXtSZWEZ>
- [25] Rohit Patel, Alexandre Rezende, and Steven McClain. 2026. GIM: Evaluating models via tasks that integrate multiple cognitive domains. *arXiv:2605.18663* [cs.AI]. <https://arxiv.org/abs/2605.18663>
- [26] Cheng Peng, Kai Zhang, Mengxian Lyu, Hongfang Liu, Lichao Sun, and Yonghui Wu. 2025. Scaling Up Biomedical Vision-Language Models: Fine-Tuning, Instruction Tuning, and Multi-Modal Learning. *arXiv preprint arXiv:2505.17436* (2025).
- [27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <https://proceedings.mlr.press/v139/radford21a.html>
- [28] Pedro Miguel Lima de Sousa Reis. 2022. Data Labeling tools for Computer Vision: a Review.
- [29] Aryan Sajith and Krishna Chaitanya Rao Kathala. 2025. Is Training Data Quality or Quantity More Impactful to Small Language Model Performance. In *Neural Computing for Advanced Applications*, Haijun Zhang, Kim Fung Tsang, Fu Lee Wang, Tianyong Hao, Zenghui Wang, Zhou Wu, Zhao Zhang, and Kevin Hung (Eds.). Springer Nature Singapore, Singapore, 75–87.
- [30] Hajar Sakai and Sarah S Lam. 2026. Large Language Models for Health Care Text Classification: Systematic Review. *JMIR AI* 5 (11 Feb 2026), e79202. doi:10.2196/79202
- [31] Silvana Secinaro, Davide Calandra, A Secinaro, V Muthurangu, and Paolo Biancone. 2021. The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making* 21, 1 (2021), 125. doi:10.1186/s12911-021-01488-9
- [32] Cora Silberschneider, Samuel S. Guimarães, Fabricio Benevenuto, and Márcio Silva. 2022. Characterizing Brazilian Political Ads on Facebook. In *Proceedings of*

- the Brazilian Symposium on Multimedia and the Web* (Curitiba, Brazil) (*WebMedia '22*). Association for Computing Machinery, New York, NY, USA, 47–54. doi:10.1145/3539637.3557935
- [33] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaeckermann, Amy Wang, Dev Dash, Jonathan H Chen, Nigam H Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera Y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S Sara Mahdavi, Joelle K Barral, Dale R Webster, Greg S Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2025. Toward expert-level medical question answering with large language models. *Nat. Med.* 31, 3 (March 2025), 943–950.
- [34] Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. 2025. The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 4195–4206. doi:10.18653/v1/2025.naacl-long.211
- [35] Haozhan Sun, Chenchu Xu, and Hanna Suominen. 2021. Analyzing the Granularity and Cost of Annotation in Clinical Sequence Labeling. *arXiv preprint arXiv:2108.09913* (2021). <https://arxiv.org/abs/2108.09913>
- [36] Petter Törnberg. 2024. Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129* (2024).
- [37] Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Cunha, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos André Gonçalves. 2019. CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 753–761. doi:10.1145/3289600.3291032
- [38] Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Gonçalves. 2020. CluHTM: Semantic Hierarchical Topic Modeling Based on CluWords. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 8138–8150. doi:10.18653/v1/2020.acl-main.724
- [39] Felipe Viegas, Washington Cunha, Christian Gomes, Leonardo Rocha, and Marcos Gonçalves. 2025. Exploiting Contextual Embeddings in Hierarchical Topic Modeling and Investigating the Limits of the Current Evaluation Metrics. *Computational Linguistics* (2025).
- [40] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. 2025. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 15134–15186. doi:10.18653/v1/2025.acl-long.736
- [41] Marco Zenone, Nora Kenworthy, and Nason Maani. 2023. The social media industry as a commercial determinant of health. *International Journal of Health Policy and Management* 12, 1 (2023), 1–4.
- [42] Marco Zenone, Jeremy Snyder, Jean-Christophe Bêlisle-Pipon, Timothy Caulfield, May CI van Schalkwyk, and Nason Maani. 2023. Advertising alternative cancer treatments and approaches on Meta social media platforms: content analysis. *JMIR Infodemiology* 3, 1 (2023), e43548.
- [43] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine* (2024), 1–13.
- [44] Zhihao Zhang, Yiran Zhang, Xiyue Zhou, Liting Huang, Imran Razzak, Preslav Nakov, and Usman Naseem. 2025. From Generation to Detection: A Multimodal Multi-Task Dataset for Benchmarking Health Misinformation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 24245–24260. doi:10.18653/v1/2025.findings-emnlp.1316
- [45] Jie Zheng et al. 2018. Clinical text annotation – what factors are associated with the cost of time?. In *AMIA Annual Symposium Proceedings*, Vol. 2018. American Medical Informatics Association, 1552–1560.
- [46] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhenhao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems* 36 (2023), 46595–46623.