Uma Análise Qualitativa da Sumarização de Diálogos de Pacientes Usando Grandes Modelos de Linguagem Aplicados a Dados Reais, Ruidosos, Informais e em Português*

Anderson A. Ferreira[1,3], Leonardo Rocha[2], Washington Cunha[1], Ana Cláudia Machado[2], João Marcos Campos[1], Gabriel Jallais[1], Adriana C. F. Viana[4], Elisa Tuler[2], Iago Araújo[1], Victor Macul[5,6], Olivio Souza Neto[5], Antônio Pereira de Souza Júnior[5], Giordano de Pinho Souza[5], Joice Marques Pallone[5], Mariana Aparecida Dumbá Soares[5], Welton Augusto Santos[5], and Marcos André Gonçalves[1]

[1]Universidade Federal de Minas Gerais, Departamento de Ciência da Computação
[2]Universidade Federal de São João del-Rei, Departamento de Ciência da Computação
[3]Universidade Federal de Ouro Preto, Departamento de Computação
[4]Universidade Federal de São João del-Rei, Departamento de Psicologia
[5]Ana Health
[6]Insper

Resumo

Este estudo avalia a capacidade de Grandes Modelos de Linguagem (*Large Language Models* - LLMs, em inglês) em resumir diálogos do mundo real, entre pacientes e a equipe de saúde de uma empresa que fornece serviços de saúde, comunicando com seus pacientes principalmente via WhatsApp. Para isso, a equipe precisa de acesso rápido às informações dos pacientes, visando fornecer respostas precisas e personalizadas. Este trabalho propõe resumir mensagens trocadas anteriomente entre pacientes e a equipe de saúde, visando construir resumos concisos, não redundantes e verdadeiros, que capturem as principais características do diálogo, apesar de trabalhar com textos ruidosos e informais (do mundo real), em um idioma sub-representado nas LLMs - o português. Para isso, foi coletado um conjunto de dados anonimizado, em português, de mensagens do WhatsApp trocadas entre pacientes e a equipe de saúde. A qualidade do diálogo foi avaliada quanto ao tamanho, legibilidade e corretude, antes da geração de resumos com duas LLMs, LLaMA 3 e Qwen 2, usando prompts específicos. Voluntários avaliaram esses resumos quanto à cobertura, relevância,

*Tadução do artigo (Ferreira et al., 2025)

redundância e veracidade, usando uma escala Likert de 5 pontos. Os resultados experimentais qualitativos e quantitativos indicam que os LLMs podem produzir resumos eficazes de diálogos entre pacientes e equipes de saúde, mesmo diante de dados de baixa qualidade e em uma linguagem sub-representada. Este é um resultado surpreendente devido ao cenário desafiador. Entre os LLMs testados, o LLaMA 3 demonstrou uma ligeira vantagem sobre o Qwen 2 sob as perspectivas de cobertura e veracidade, entre os métodos avaliados. Os resultados demonstram o potencial para construir serviços práticos do mundo real, podendo auxiliar profissionais de saúde a responder a mensagens de pacientes com agilidade, clareza e coesão, aumentando a eficiência da comunicação e a satisfação do paciente. Por fim, a abordagem aqui proposta pode melhorar significativamente o cenário da comunicação online em saúde, particularmente em ambientes com recursos limitados como o Brasil, onde o acesso à atenção primária é limitado.

1 Introdução

O uso crescente de prontuários eletrônicos, bem como de plataformas de mensagens (instantâneas), como WhatsApp e Messenger, expandiu o uso de portais digitais de saúde para a troca de mensagens entre pacientes e equipes de saúde, aumentando consideravelmente o volume de mensagens trocadas nesses ambientes (Liu et al., 2024). Com isso, gerenciar esse volume crescente de mensagens é um desafio. De fato, estima-se que médicos de atenção primária gastem cerca de 1,5 hora por dia analisando aproximadamente 150 mensagens de pacientes (Liu et al., 2024). Para responder de forma eficaz, os profissionais de saúde devem compreender o contexto da mensagem, revisando mensagens anteriores do mesmo paciente, bem como quaisquer outras informações relevantes disponíveis sobre o paciente.

No Brasil, onde 34% da população não tem acesso à atenção primária, estima-se que atingir 100% de cobertura exigiria 236.900 profissionais de saúde, a um custo de R$ 22,9 bilhões/ano (Hone et al., 2017). Uma abordagem para escalar esse serviço seria por meio de um modelo de Atenção Primária Digital, como o oferecido pela Ana Health (*https://www.anahealth.com.br/*), uma empresa de saúde digital que fornece serviços de saúde, comunicando-se principalmente via WhatsApp, para poder atender o maior número de pacientes. Os profissionais de saúde da Ana Health interagem principalmente com a população por meio de um sistema de mensagens específico — o WhatsApp. Nesse modelo de interação, o desafio de processar e compreender um grande volume de mensagens permanece, pois é essencial responder a cada paciente de forma personalizada.

De uma forma mais detalhada, na Ana Health, quando um membro da equipe de saúde recebe uma nova mensagem de um paciente, para responder adequadamente, ele deve compreender todo o contexto da mensagem, revisando mensagens anteriores trocadas com o paciente, além de entender a mensagem atual. Somente depois disso, o membro

da equipe de saúde pode elaborar uma resposta que permita manter o engajamento do paciente. Para que o paciente receba a resposta de forma eficiente e eficaz, é essencial coletar rapidamente informações contextuais, que possibilitem fornecer respostas precisas, atualizadas e personalizadas.

Uma abordagem para auxiliar a equipe de saúde a compreender o contexto das mensagens do paciente, que é analisado e proposto neste trabalho, é aplicar técnicas de sumarização a mensagens anteriores trocadas entre o paciente e a equipe de saúde. Essa sumarização, se bem-sucedida, também pode auxiliar outros serviços que a equipe de saúde oferece ao paciente, como psicoterapia, monitoramento de doenças crônicas e prescrição de medicamentos.

Alguns trabalhos têm desenvolvido e avaliado técnicas de sumarização (El-Kassas et al., 2021; Keszthelyi et al., 2023; Zhang; Yu; Zhang, 2025), incluindo a sumarização de dados clínicos (Keszthelyi et al., 2023), explorando técnicas que variam desde abordagens baseadas em regras até o uso de Grandes Modelos de Linguagem (*Large Language Models* - LLMs, em inglês) (Minaee et al., 2024). No entanto, essas técnicas raramente são avaliadas em ambientes reais, usando dados reais (Laskar et al., 2024). Este trabalho busca preencher essa lacuna na literatura. Ou seja, busca avaliar a capacidade dos LLMs em gerar resumos úteis, que possam auxiliar uma equipe de saúde a compreender o contexto de um paciente, usando dados de saúde reais e anonimizados, além de conduzir uma avaliação completa desses resumos com o apoio de especialistas, particularmente voluntários de uma equipe de saúde.

Avanços recentes em LLMs baseados em transformadores têm auxiliado em diversas tarefas relacionadas ao processamento de linguagem natural, incluindo sumarização (Minaee et al., 2024). Em (Yang et al., 2023), os autores afirmam que os LLMs podem auxiliar em diversas tarefas na área de saúde, como pré-consulta, diagnóstico e gerenciamento, com desenvolvimento e supervisão adequados. Apesar disso, algumas preocupações e desafios ainda permanecem no uso de LLMs na área de saúde, incluindo (Dave; Athaluri; Singh, 2023; Wang et al., 2024; Yang et al., 2023): (1) privacidade de dados – garantir que dados reais de pacientes não sejam vazados e usados pelos proprietários dos LLMs para outros fins; (2) credibilidade e correção das informações – os LLMs podem perpetuar informações incorretas ou produzir "alucinações"; e (3) dados tendenciosos – os LLMs são comumente treinados com dados diversos que podem ser tendenciosos e, consequentemente, gerar respostas tendenciosas.

Voltando à tarefa em foco, neste trabalho, é utilizada a sumarização -— um processo de redução de textos longos para uma versão mais curta, preservando informações importantes para a compreensão do texto original (Mani; Maybury, 1999; Zhang; Yu; Zhang, 2025). O objetivo é avaliar a capacidade de LLMs em resumir diálogos (completos) do WhatsApp entre pacientes e uma equipe de saúde. Idealmente, esses resumos devem capturar as principais características dos diálogos, ser concisos, não redundantes e conter informações

verdadeiras.

Especificamente, é proposto avaliar LLMs de código aberto e disponíveis publicamente sobre a tarefa de resumir mensagens de pacientes, auxiliando profissionais de saúde com informações concisas, orientadas ao paciente e de fácil leitura, com a finalidade de auxiliar a elaboração de respostas a esses pacientes. Além disso, tal resumo deve permitir que a equipe de saúde escale o gerenciamento do cuidado com os pacientes de forma personalizada, adaptando abordagens diversas às necessidades dos pacientes, servindo como um passo fundamental para o desenvolvimento de um agente digital de inteligência artificial para a atenção primária. Para avaliar a viabilidade de tal abordagem, foram utilizados dados de ambientes reais, incluindo interações reais (anonimizadas) entre pacientes e equipes de saúde, e foi conduzida uma avaliação desses resumos com o apoio de voluntários de uma equipe de saúde real.

Alguns estudos foram conduzidos sobre sumarização de dados clínicos. Em particular, em (Keszthelyi et al., 2023), os autores revisaram publicações sobre sumarização de dados no contexto clínico. Os dados sumarizados incluem principalmente prontuários eletrônicos de pacientes. Os autores também discutem a importância de tais sumarizações para auxiliar tarefas relacionadas a unidades de terapia intensiva, cirurgia, diagnóstico, cuidados hospitalares, monitoramento de doenças crônicas, oncologia, prescrição de medicamentos, psicoterapia, etc. Este trabalho difere de (Keszthelyi et al., 2023) principalmente em dois aspectos: (1) fonte de dados — este trabalho explora dados provenientes de um sistema de mensagens instantâneas, particularmente mensagens de diálogos entre pacientes e uma equipe de saúde; tais mensagens geralmente são ruidosas, podendo impactar a qualidade da sumarizção; e (2) domínio de aplicação — o principal objetivo é auxiliar uma equipe de saúde a responder a novas mensagens de pacientes, garantindo respostas personalizadas, humanísticas e contextualmente apropriadas.

Assim, a principal e inovadora contribuição deste trabalho é a avaliação da capacidade de dois grandes modelos de linguagem, LLaMA 3 e Qwen 2, em resumir diálogos reais entre pacientes e equipes de saúde, via WhatsApp, com base em dados reais, ruidosos e informais, em Português. O estudo destaca o potencial dos LLMs em aprimorar a comunicação digital na área de saúde, particularmente em cenários com recursos limitados como o Brasil, onde o acesso à atenção primária é limitado.

2 Métodos

2.1 Conjunto de dados e pré-processamento

Esta pesquisa foi conduzida no Laboratório de Banco de Dados da Universidade Federal de Minas Gerais. O projeto de pesquisa, incluindo os dados e a participação de voluntários em avaliações qualitativas, foi revisado e aprovado pelo Comitê de Ética da Universidade

Federal de Minas Gerais sob o registro CAAE 80632524.4.0000.5149, em conformidade com a Resolução CNS 446/12.

Este estudo reuniu mensagens de diálogos entre pacientes, equipe de saúde, via WhatsApp, de 27 de outubro de 2021 a 8 de janeiro de 2024, por meio de soluções da Ana Health baseadas em nuvem. Este conjunto de dados contém 207.040 mensagens escritas em português. Todos os experimentos foram conduzidos de acordo com as diretrizes e normas pertinentes e aprovados pelo Comitê de Ética da Universidade Federal de Minas Gerais. Ressaltamos que, o consentimento para o uso do conjunto de dados foi obtido do responsável legal.

A Ana Health é uma empresa que oferece um serviço digital de atenção primária à saúde física e mental das pessoas. Por meio de um plano de assinatura, os associados (pacientes) têm acesso ilimitado à equipe multidisciplinar da Ana Health, incluindo clínicos gerais, psicólogos, enfermeiros e gerontólogos, que os orientam proativamente em uma jornada de saúde estruturada, complementada por suporte 24 horas por dia, 7 dias por semana, por meio de mensagens de texto e chamadas telefônicas ou de vídeo. A empresa visa escalar a atenção primária por meio de uma experiência prioritáriamente via WhatsApp, impulsionada por um sistema de gestão de cuidados personalizado, alimentado por inteligência artificial. A Ana Health disponibilizou essas mensagens de forma anônimizada, garantindo que todas as informações identificáveis, como nomes, e-mails, números de telefone e URLs, fossem removidas e substituídas por um código exclusivo para proteger a privacidade dos dados.

Em relação ao pré-processamento, foram removidos caracteres especiais (tabulação, marcador de nova linha e espaços em branco duplicados) e filtradas mensagens modelo — mensagens padronizadas da Ana Health para seus pacientes — resultando em 202.326 mensagens e 1.863 diálogos.

2.2 Análise da qualidade das mensagens

Antes de resumir os diálogos, a qualidade textual das mensagens foi analisada, para compreender os limites dos dados e o impacto potencial da qualidade dos dados nos resultados finais. Textos de baixa qualidade podem levar os LLMs a gerar resumos incoerentes ou imprecisos, ou resumos que não contêm as informações principais do texto. Além disso, erros ortográficos ou gramaticais podem ser interpretados pelos LLMs como palavras ou conceitos desconhecidos ou inapropriados. A análise deste trabalho foi baseada na metodologia proposta por Dalip et al. (2009). Foram analisadas três dimensões da qualidade textual: (1) tamanho, (2) legibilidade e (3) corretude.

Tamanho

Número total de caracteres, palavras e frases. De acordo com Dalip et al. (2009), mensagens de alta qualidade geralmente não são nem muito curtas nem muito longas. Também definimos mensagens longas como aquelas que excedem o comprimento médio em pelo menos dez palavras, e mensagens curtas como aquelas que contêm cinco palavras a menos que o comprimento médio da mensagem ou menos, semelhante ao que foi feito em (Dalip et al., 2009).

Legibilidade

Este trabalho utiliza a identificação do nível de escolaridade Flesch-Kincaid (Kincaid et al., 1975), que gera uma nota geral avaliando a complexidade de um texto, de acordo com a Equação 1.

$$0,39 \times \frac{total\_de\_palavras}{total\_de\_sentenças} + 11,8 \times \frac{total\_de\_sílabas}{total\_de\_palavras} - 15,59 \quad (1)$$

Como resultado, é obtida uma pontuação que relaciona com os níveis de escolaridade norte americano. O nível de escolaridade Flesch-Kincaid considera frases com muitas palavras ou palavras contendo muitas sílabas mais difíceis de compreender. O nível de escolaridade Flesch-Kincaid é uma métrica de legibilidade bem estabelecida, que estima o nível mínimo de escolaridade necessário para compreender um determinado texto. Na análise deste trabalho, não foi utilizado o valor Flesch-Kincaid para avaliar se os textos eram bem escritos. Em vez disso, o objetivo foi realizar uma análise comparativa dos níveis de legibilidade de textos escritos por pacientes versus aqueles escritos pela equipe de saúde. Embora o nível de escolaridade Flesch-Kincaid tenha sido originalmente desenvolvido para indicar o nível de escolaridade necessário para a compreensão de um documento escrito em inglês, existem estudos (Martins et al., 1996; Moreno et al., 2022) que o adaptam para avaliar textos em português. Esses trabalhos também mostram uma alta correlação entre os valores obtidos pelas fórmulas original e a adaptada. Assim, pode-se usar a fórmula original do nível de escolaridade Flesh-Kincaid para comparar documentos escritos na mesma língua, como feito neste trabalho.

Corretude

Calcula-se a proporção de palavras nas mensagens que estão presentes em um dicionário da lingua portuguesa predefinido. Foi utilizado o dicionário do pacote *br.ispell* (*https://github.com/fititnt/br.ispell-dicionario-portugues-brasileiro*).

2.3 Sumarização dos diálogos

Como mencionado, o principal objetivo deste trabalho é avaliar a capacidade dos LLMs em resumir diálogos entre pacientes e equipe de saúde. Idealmente, tais resumos devem capturar os principais aspectos dos diálogos e não incluir informações incorretas. Para fins de reprodutibilidade e transparência, este estudo foca no uso de LLMs de código aberto, disponíveis publicamente, que podem ser executados localmente, garantindo a privacidade dos dados, já que não é enviado dados privados de pacientes a LLMs remotos e proprietários, como o GPT. Nos experimentos, avaliou-se o Qwen 2 (Qwen2-7B-Instruct) e o LLaMA 3 (Meta-Llama-3-8B-Instruct), como LLMs para resumir os diálogos. O LLaMA 3 e Qwen 2 foram selecionados com base em várias considerações importantes:

1. Disponibilidade de Código Aberto e Privacidade dos Dados: Ambos os modelos são de código aberto, permitindo implantação local sem depender de infraestrutura de nuvem externa. Este aspecto é crucial para a aplicação, pois permite que o sistema opere sem transmitir dados sensíveis do usuário via Internet, mitigando assim potenciais riscos relacionados à privacidade e segurança dos dados, especialmente relevantes ao considerar a futura adoção em nível de produção pela empresa parceira.

2. Desempenho empírico: LLaMA 3 e Qwen 2 demonstraram desempenho de ponta em uma variedade de tarefas de processamento de linguagem natural, incluindo perguntas e respostas, classificação, geração de código, reescrita de texto e raciocínio. Benchmarks e avaliações públicas (por exemplo, o blog Meta AI sobre LLaMA 3 e o Hugging Face model card para Qwen2-7B) comprovam a superioridade desses modelos em comparação a outras alternativas de código aberto.

3. Tamanho do modelo e eficiência computacional: Foram utilizadas as versões de 8 bilhões de parâmetros do LLaMA 3 e de 7 bilhões de parâmetros do Qwen 2. Esses modelos alcançam um equilíbrio entre desempenho e custo computacional, permitindo executá-los em hardware local padrão, sem exigir infraestrutura especializada ou de alto custo.

Para o processo de sumarização, todas as mensagens de um diálogo entre paciente e equipe de saúde foram agrupadas, para gerar um único documento (possivelmente um documento longo). Mais formalmente, para cada paciente p_i pertencente a um conjunto de pacientes $\{p_1, p_2, \cdots p_n\}$, produziu-se um documento d_i com todas as mensagens relacionadas a p_i. Um método de sumarização recebe um documento d_i para p_i como entrada e deve produzir um resumo s_i capturando os principais aspectos e tópicos discutidos em d_i.

Para usar LLMs como método de sumarização, foi desenvolvido um *prompt*, ilustrado na Figura 1. Este *prompt* contém instruções para executar a tarefa de sumarização. O texto a ser processado por um LLM deve ser *tokenizado*, ou seja, dividido em unidades

menores de subpalavras conhecidas como *tokens*, resultando em uma sequência de tokens. Cada subpalavra é identificada numericamente. O LLM processa a sequência de tokens e gera uma sequência diferente que dará origem ao respectivo resumo. A Figura 2 ilustra o processo completo de sumarização.

Figura 1 – Exemplo de *prompt*

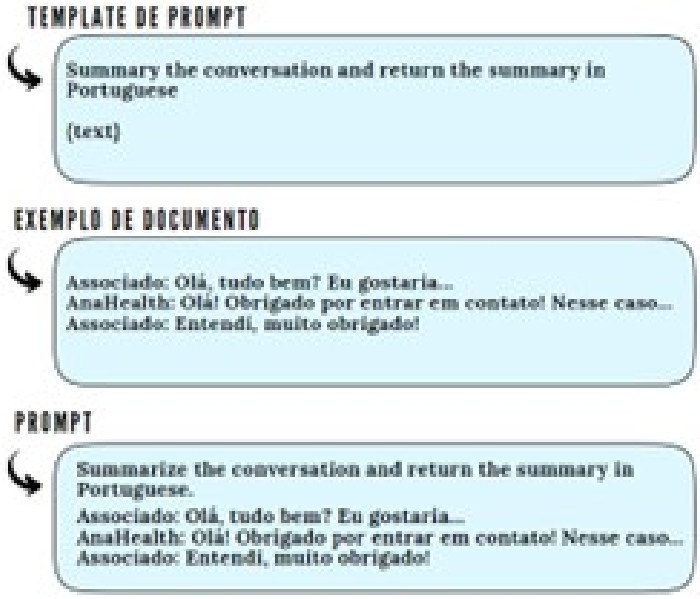
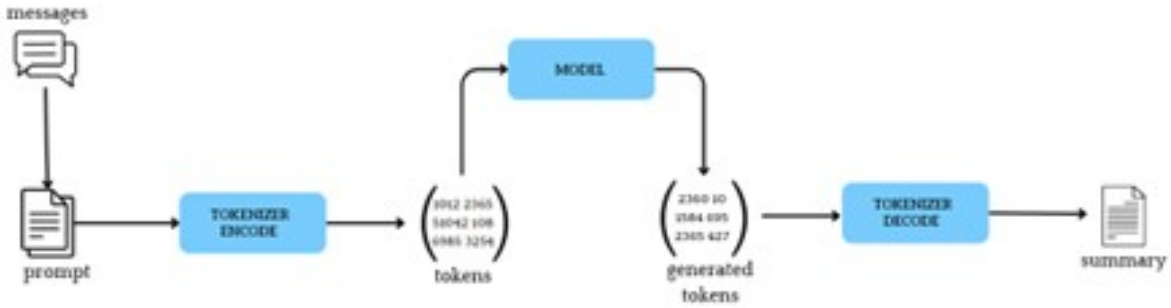


Figura 2 – Processo de sumarização com LLM



Os LLMs podem receber apenas um número limitado de tokens como entrada. Neste trabalho, devido à ordem cronológica das mensagens, optamos por usar os últimos 5.000 tokens de um diálogo como entrada para os LLMs, pois as últimas mensagens do diálogo podem ser mais importantes, contendo informações mais atuais, para responder a uma recente mensagem de um paciente.

2.4 Avaliação dos resumos

Para avaliar os resumos gerados pelos LLMs, é realizado um teste A/B. A avaliação dos resumos gerados foi feita com a participação de voluntários da equipe de saúde da Ana Health. É importante ressaltar que os participantes que avaliaram os resumos não participaram dos diálogos com os pacientes, para evitar qualquer viés vindo de conhecimentos prévios sobre os pacientes.

Para avaliar os resumos gerados, não foram utilizadas métricas tradicionais de avaliação de textos, como ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), NIST (Doddington, 2002) e METEOR (Banerjee; Lavie, 2005), que são baseadas na sobreposição de textos, ou BERTScore (Zhang et al., 2019) e BARTScore (Yuan; Neubig; Liu, 2021), que são métricas baseadas em similaridade, pois essas métricas requerem informações sobre um texto de referência para realizar a comparação, o que não está disponível em nosso conjunto de dados. Em outras palavras, nosso conjunto de dados não tem informação sobre os "resumos perfeitos". Alguns trabalhos recentes também analisam o uso de LLMs como avaliadores de resumos (Gao et al., 2023; Jain et al., 2023; Wu et al., 2023). No entanto, o objetivo não é analisar a capacidade dos LLMs em avaliar a qualidade dos resumos. Em vez disso, este trabalho foca em avaliar seu desempenho na geração de resumos úteis sobre pacientes, diante de dados ruidosos e informais, com o objetivo de auxiliar o profissional de saúde a responder adequadamente a mensagens de pacientes. Assim, neste trabalho, é conduzida uma avaliação por humanos, em particular, potenciais usuários reais dos resumos gerados. Isso constitui uma das principais contribuições deste trabalho.

Critério de avaliação dos resumos

Este trabalho elaborou um questionário, no qual cada participante deveria ler as mensagens de um diálogo original (conversa) na íntegra e, em seguida, classificar cada resumo gerado utilizando uma escala Likert de 5 pontos (1 – discordo totalmente, 5 – concordo totalmente) (Gao et al., 2023), considerando quatro perspectivas:

- Cobertura: O resumo abrange todos os aspectos importantes e relevantes da conversa completa;
- Relevância: Todos os aspectos do resumo são relevantes para o contexto das mensagens, ou seja, o resumo não contém partes irrelevantes;
- Redundância: O resumo não é redundante, ou seja, não contém informações repetidas;
- Veracidade: O resumo é verdadeiro (não contém informações incorretas).

Após revisar a literatura (El-Kassas et al., 2021; Fabbri et al., 2021) e discutir as métricas com a equipe de saúde da Ana Health, decidiu-se focar a avaliação dos resumos sobre essas perspectivas. O Apêndice **??** contém um exemplo do questionário.

Participantes

Este trabalho contou com a participação de cinco médicos e 19 psicólogos para avaliar os resumos. Foram selecionados 24 profissionais de saúde responsáveis por acolher os pacientes, realizar a triagem com base nas informações relatadas pelos pacientes e manter um entendimento preciso do histórico de cada paciente e das interações mais recentes com a empresa. Também é importante destacar que a literatura indica que testes de usabilidade com mais de 15 usuários costumam ser suficientes para identificar a grande maioria — senão todos — dos principais problemas (Nielsen, 2000). A pesquisa foi distribuída entre os participantes para garantir uma representação equilibrada dos tipos de profissionais. Como informação adicional, os participantes são predominantemente mulheres, com aproximadamente 22% sendo homens. Cada resumo foi avaliado por duas mulheres e um homem, ou por três mulheres.

Protocolo de avaliação

Nesta pesquisa, foi aplicado o mesmo protocolo de avaliação a todos os participantes (externos e internos) para minimizar o viés em nossos resultados. Estimou-se que, para manter um tempo de avaliação em torno de 15 minutos, cada participante deveria avaliar dois resumos para evitar fadiga e garantir a participação efetiva durante todo o estudo, incluindo um psicólogo externo. Vários estudos (Dillman; Smyth; Christian, 2014; Revilla; Ochoa, 2017), na área de metodologias de pesquisa, indicam que pesquisas com duração inferior a 15 minutos tendem a ter uma taxa de resposta mais alta, menos desistências no meio do preenchimento do questionário e respostas de melhor qualidade. Para cada resumo, cada participante deve ler primeiramente todo o diálogo, depois o resumo correspondente e, por fim, avaliar as quatro perspectivas com base na escala Likert.

Antes de selecionar aleatoriamente os diálogos a serem resumidos neste trabalho, foram selecionados os diálogos com uma contagem de tokens entre 1.000 e 5.000. Essa etapa garante que os diálogos não sejam truncados pelo LLM nem sejam muito breves, a ponto de conter informações insuficientes para a geração de um resumo. Em seguida, foram selecionados aleatoriamente oito diálogos completos e fornecidos como entrada aos dois LLMs -— LLaMA 3 e Qwen 2. Consequentemente, foram obtidos dezesseis resumos para avaliação, sendo cada resumo avaliado por três participantes diferentes.

Análise dos questionários

Para analisar as respostas dos questionários, primeiramente foram analisadas as distribuições de respostas em cada perspectiva, visando obter uma compreensão inicial do comportamento das avaliações dos participantes sobre cada LLM. Em seguida, foi utilizada a escala numérica (1 - discordo totalmente – 5 - concordo totalmente) para analisar os

valores médios e seus desvios-padrão, a fim de obter uma interpretação quantitativa da eficácia dos LLMs.

Também calculou-se o coeficiente Kappa-Cohen (k) para medir a concordância entre participantes. Este coeficiênte é útil para avaliar a consistência das respostas obtidas. Esta métrica foi aplicada exclusivamente sobre avaliações na mesma perspectiva sobre o mesmo resumo e por dois participantes. Após calcular a concordância para todos os pares possíveis de respostas, calcula-se a média para determinar um valor de concordância geral.

$$k = \frac{p_o - p_e}{1 - p_e} \tag{2}$$

sendo que p_o representa a taxa de concordância observada, refletindo a probabilidade empírica de concordância nas respostas fornecidas para qualquer afirmação, e p_e representa a concordância esperada, assumindo que os avaliadores respondem aleatoriamente. Ao incorporar p_e, considera-se que a possibilidade de concordância ocorrer por acaso.

Depois, é realizada uma análise qualitativa dos resumos gerados em neste estudo, utilizando a perspectiva de um profissional focado no acolhimento humanizado de pacientes no contexto do processo psicoterapêutico.

Privacidade dos dados

Este estudo foi conduzido em acordo com os mais altos padrões éticos e recebeu aprovação prévia de um comitê de ética institucional independente. Foi dada atenção específica às questões de anonimização de dados, preservação da privacidade e às implicações éticas do processamento de comunicações relacionadas aos pacientes.

O conjunto de dados utilizado em neste estudo foi anonimizado antes do nosso acesso, com todas as informações identificáveis removidas pela empresa parceira, em conformidade com suas políticas internas de governança de dados. Os autores deste trabalho não realizaram nenhuma coleta de dados por conta própria. Em vez disso, receberam os dados pré-processados de forma anônimizada, sem acesso a quaisquer identificadores pessoais, como nomes, informações de contato ou números de identificação de saúde.

Além disso, todo o processamento dos dados, incluindo a tarefa de sumarização, foi realizado localmente, utilizando infraestrutura local. Em nenhum momento, os dados ou resultados derivados foram transmitidos para serviços baseados em nuvem ou servidores externos. Essa abordagem garante a eliminação do risco de exposição de dados por meio de canais de comunicação inseguros.

Para reforçar nosso compromisso com a confidencialidade, todos os membros da equipe de pesquisa assinaram um Acordo de Confidencialidade juridicamente vinculativo com a entidade fornecedora dos dados, garantindo que nenhuma parte dos dados ou resultados possa ser usada ou divulgada fora do escopo do estudo aprovado.

Em relação à implantação operacional do sistema, é importante esclarecer que somente profissionais de saúde autorizados, que já tenham acesso às comunicações originais dos pacientes dentro do fluxo de trabalho clínico da empresa, poderão acessar os resumos gerados pelo sistema. Esses profissionais estão, por definição, vinculados à confidencialidade profissional e a obrigações éticas, e já possuem acesso a informações mais detalhadas e potencialmente identificáveis do que as presentes nos resumos. Portanto, o componente de sumarização não introduz risco adicional à privacidade, nem expande o acesso aos dados, além do que já está sancionado pela empresa.

Foram aplicadas técnicas de anonimização de última geração, removendo identificadores diretos e garantindo que todo o trabalho fosse conduzido sob rigorosas salvaguardas legais e éticas. Essas medidas mitigam significativamente o risco de desanonimização e são consistentes com as práticas éticas de pesquisa e os regulamentos de proteção de dados, como as leis nacionais de proteção de dados aplicáveis a este contexto.

3 Resultados

3.1 Avaliação da qualidade das mensagens

Nesta seção, é discutida a qualidade das mensagens considerando as três dimensões: tamanho, legibilidade e corretude.

Tamanho

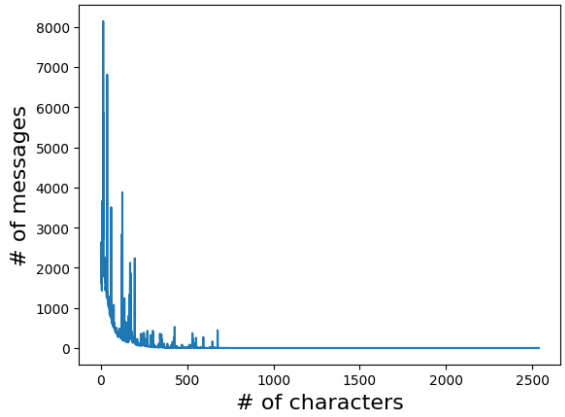
A Figura 3 mostra a distribuição do tamanho das mensagens em relação à quantidade de palavras, caracteres e frases, considerando todas as mensagens trocadas entre pacientes e a equipe de saúde, enquanto a Tabela 1 mostra os valores médios, máximos e mínimos de caracteres, palavras e frases -– juntamente com seus limites de quartis — considerando todas as mensagens. Até 75% das mensagens têm até 135 caracteres, 25 palavras e 4 frases, respectivamente.

Tabela 1 – Tamanho - média (desvio padrão), valores mínimo e máximo e limites dos quartis (Q1, Q2 e Q3) de 202.326 mensagens. Até 75% das mensagens têm 135 caracteres, 25 palavras e 4 frases.

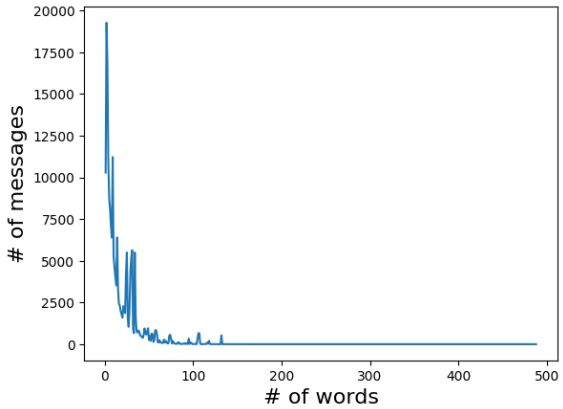
	Média (DP)	Mínimo	Máximo	Q1	Q2	Q3
Caracteres	91,45 (109,49)	1	2.538	17	48	135
Palavras	17,47 (19,95)	1	488	4	10	25
Frases	2,71 (2,01)	1	24	1	2	4

Foram encontradas 47.443 (23,45%) mensagens longas e 112.971 (55,84%) mensagens curtas. Ou seja, a maioria das mensagens são curtas, enquanto quase um quarto são longas.

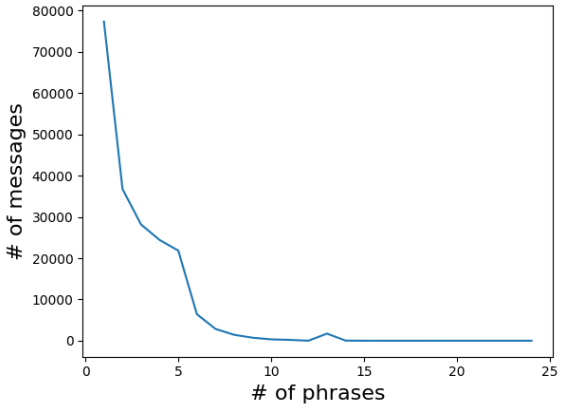
Figura 3 – Distribuição da quantidade de (a) caracteres, (b) palavras e (c) frases em todas as mensagens (202.326 mensagens) trocadas entre pacientes e a equipe de saúde. A maioria das mensagens contém poucos caracteres, palavras e frases.



(a) Número total de caracteres considerando todas as mensagens



(b) Número total de palavras considerando todas as mensagens



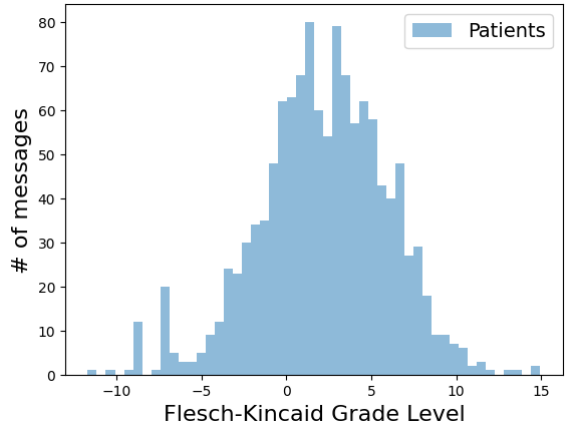
(c) Número total de frases considerando todas as mensagens

Legibilidade

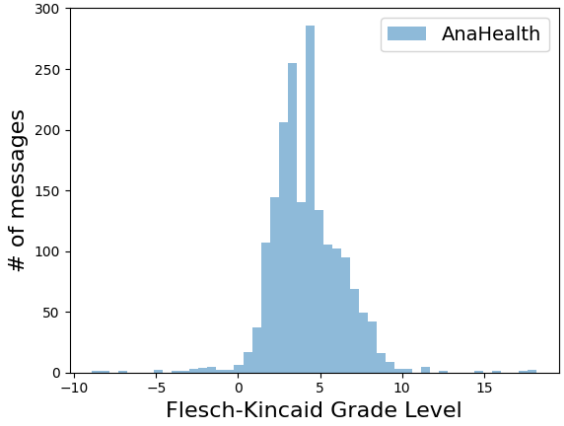
Foi utilizado o nível de escolaridade Flesch-Kincaid para a análise de legibilidade. Antes de aplicar o nível de escolaridade Flesch-Kincaid, todas as mensagens de e para o mesmo paciente foram agrupadas, resultando em três tipos de "documentos": (i) todas as mensagens enviadas pela equipe de saúde para um paciente p_i; (ii) todas as mensagens enviadas por p_i para a equipe de saúde; e (iii) a união de ambas, ou seja, todas as mensagens de/ou para cada paciente (todas as mensagens por paciente). Conforme discutido acima, foram separadas as análises de acordo com a direção do fluxo de mensagens. A Figura 4 e a Tabela 2 mostram os resultados.

Comparando os níveis de escolaridade Flesch-Kincaid dos grupos de mensagens enviadas pela equipe de saúde com os dos grupos de mensagens enviadas pelos pacientes, observam-se valores médios e limites de quartis mais elevados para as mensagens enviadas pela equipe de saúde. Isso indica que, em média, os textos produzidos pela equipe de saúde

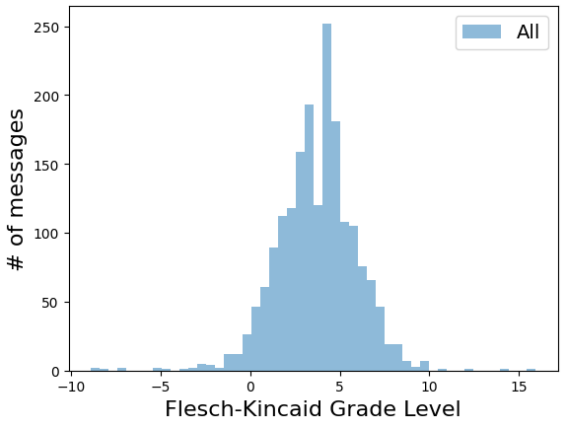
Figura 4 – Distribuição do nível de escolaridade Flesch-Kincaid para (a) grupos de mensagens enviadas por cada paciente à equipe de saúde, (b) grupos de mensagens enviadas pela equipe de saúde a cada paciente e (c) grupos de mensagens enviadas ou recebidas por cada paciente. As mensagens enviadas pelos pacientes têm uma gama mais ampla de níveis de escolaridade, o que significa que as mensagens dos pacientes estão distribuídas por um intervalo de níveis de escolaridade mais amplo em comparação às mensagens enviadas pela equipe de saúde.



(a) Nível de escolaridade Flesch-Kincaid sobre os grupos de mensagens enviadas pelo mesmo paciente



(b) Nível de escolaridade Flesch-Kincaid sobre os grupos de mensagens enviadas para o mesmo paciente



(c) Nível de escolaridade Flesch-Kincaid sobre os grupos de mensagens enviadas e recebidas pelo mesmo paciente

são mais elaborados e menos simplistas do que aqueles produzidos pelos pacientes (exigem um nível de escolaridade mais elevado para serem lidos). A diferença é notável e deve ser considerada no desenvolvimento de futuras metodologias que se baseiem nesse tipo de dado. A distribuição dos níveis de escolaridade Flesch-Kincaid das mensagens enviadas pelos pacientes apresenta uma amplitude maior de valores, o que significa que as mensagens dos pacientes apresentam maior variabilidade em termos de nível de escolaridade.

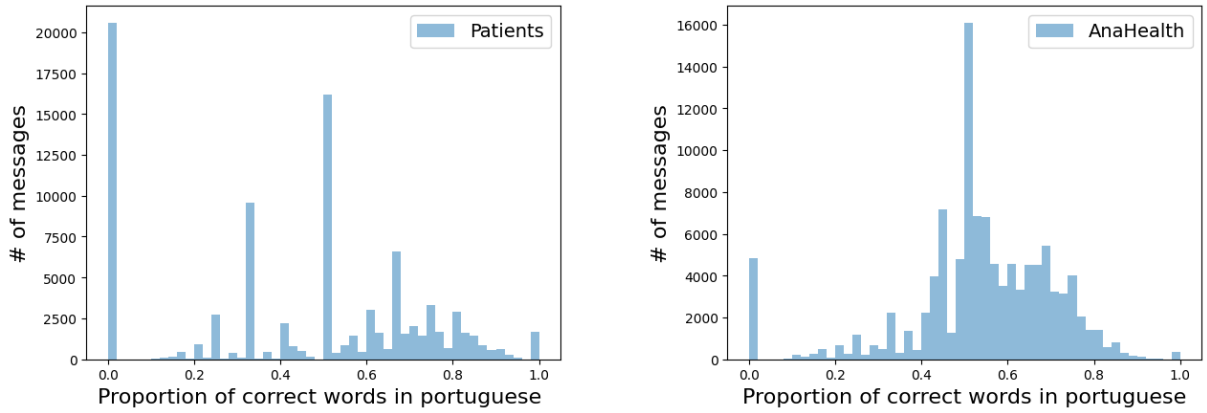
Tabela 2 – Legibilidade - número total de grupos de mensagens, média (desvio padrão), níveis mínimo e máximo de nível de escolaridade Flesch-Kincaid e limites de quartis (Q1, Q2 e Q3). Valores médios e limites de quartis altos para os grupos de mensagens enviadas pela equipe de saúde indicam que as mensagens enviadas pela equipe de saúde estão relacionadas a um nível de escolaridade mais alto.

Critério de agrupamento	# de grupos	Nível de escolaridade Flesch-Kincaid					
		Média (DP)	Mínimo	Máximo	Q1	Q2	Q3
Enviada ou recebida pelo mesmo paciente	1.863	3,72 (2,20)	-8,91	15,95	2,36	3,88	4,99
Recebida para o mesmo paciente	1.859	4,18 (2,16)	-8,91	18,20	2,82	4,08	5,35
Enviada pelo mesmo paciente	1.224	2,18 (3,85)	-11,69	14,95	-0,18	2,32	4,82

Corretude

Para esta análise, foi utilizado um dicionário da língua portuguesa, obtido do site *br.ispell*, para comparar a proporção de palavras escritas corretamente nas mensagens enviadas pelos pacientes com as enviadas pela equipe de saúde. Os resultados são mostrados na Figura 5 e na Tabela 3.

Figura 5 – Distribuição de palavras escritas corretamente em mensagens enviadas por (a) pacientes e (b) equipe de saúde. As mensagens enviadas pela equipe de saúde apresentam maior nível de corretude em comparação com as enviadas pelos pacientes. Total de 202.326 mensagens, sendo 111.011 mensagens da equipe de saúde e 91.315 dos pacientes.



(a) Paciente

(b) Equipe de saúde

Observe que as mensagens enviadas pela equipe de saúde apresentam um nível de corretude maior em comparação com as enviadas pelos pacientes. As mensagens dos

Tabela 3 – Corretude - média (desvio padrão), valores mínimo e máximo e limites de quartis (Q1, Q2 e Q3). Total de 202.326 mensagens, sendo 111.011 mensagens da equipe de saúde e 91.315 mensagens de pacientes.

Origem	Média (DP)	Mínimo	Máximo	Q1	Q2	Q3
Todas as mensagens	0,49 (0,24)	0,00	1,00	0,40	0,52	0,67
Mensagens da equipe de saúde	0,54 (0,18)	0,00	1,00	0,47	0,54	0,66
Mensagens de pacientes	0,44 (0,29)	0,00	1,00	0,25	0,50	0,67

pacientes apresentam um nível de corretude muito baixo, indicando baixa qualidade de escrita, com muitos erros de digitação, erros de ortografia, abreviações, palavras desconhecidas, etc.

3.2 Avaliação da tarefa de sumarização

Nesta seção, é discutida a avaliação da tarefa de sumarização. Antes de descrever os resultados, são mostradas algumas estatísticas sobre o tamanho (contagem de palavras) da nossa amostra de diálogos e de seus resumos correspondentes gerados por LLaMA 3 e Qwen 2. A Tabela 4 mostra a quantidade de palavras nos diálogos originais e em seus resumos correspondentes gerados por LLaMA 3 e Qwen 2. Vale ressaltar que não há uma correlação clara entre o tamanho dos diálogos de entrada e o tamanho dos resumos gerados pelos modelos de linguagem. Em média, os resumos gerados por LLaMA 3 e Qwen 2 contêm aproximadamente 137 e 147 palavras, respectivamente, com desvios-padrão de 59 e 41 palavras. Isso indica que os modelos produzem resumos com tamanhos em torno de 150 palavras, independentemente do tamanho do diálogo original.

Tabela 4 – Quantidade de palavras dos diálogos e correspondentes resumos gerados por LLaMA 3 e Qwen 2.

	# de palavras		
Diálogo	Diálogo	Resumo do LLaMA	Resumo do Qwen
1	967	102	195
2	2069	208	202
3	1254	85	164
4	1458	220	167
5	1422	106	138
6	2300	67	124
7	2767	122	107
8	2659	189	85

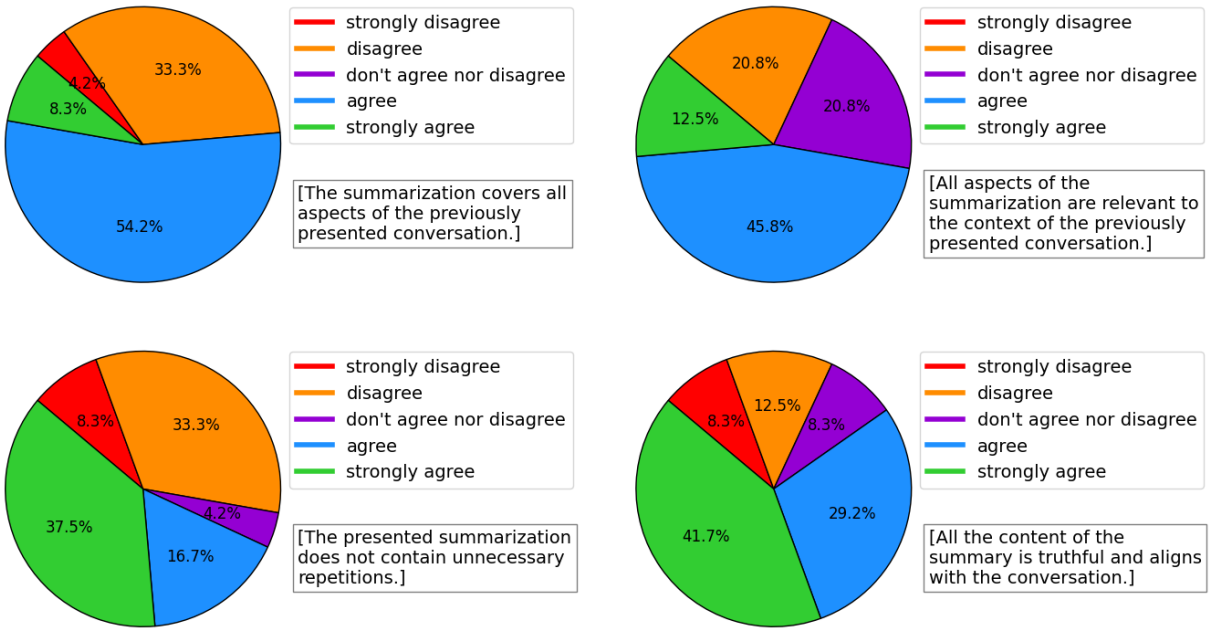
A avaliação da tarefa de sumarização foi realizada por meio de uma pesquisa com 24 participantes, incluindo cinco médicos e 18 psicólogos da equipe de saúde da Ana Health, além de um psicólogo externo. Para analisar os dados da pesquisa, foram utilizadas duas abordagens. A primeira analisa as distribuições das pontuações que os participantes

atribuíram aos resumos avaliados. E a segunda emprega métricas estatísticas para obter uma interpretação quantitativa da eficácia dos LLMs.

Distribuição das respostas por perspectiva

Cada perspectiva foi analisade individualmente, permitindo identificar padrões de respostas e comparar diretamente os LLMs. A Figura 6 ilustra a distribuição de respostas para cada perspectiva ao utilizar o LLaMA 3 na tarefa de sumarização.

Figura 6 – Distribuição das pontuações para cada perspectiva (24 avaliações por perspectiva), avaliando resumos gerados pelo LLaMA 3. As respostas positivas -— "concordo" e "concordo totalmente" -— representam mais de 50% do total de respostas.

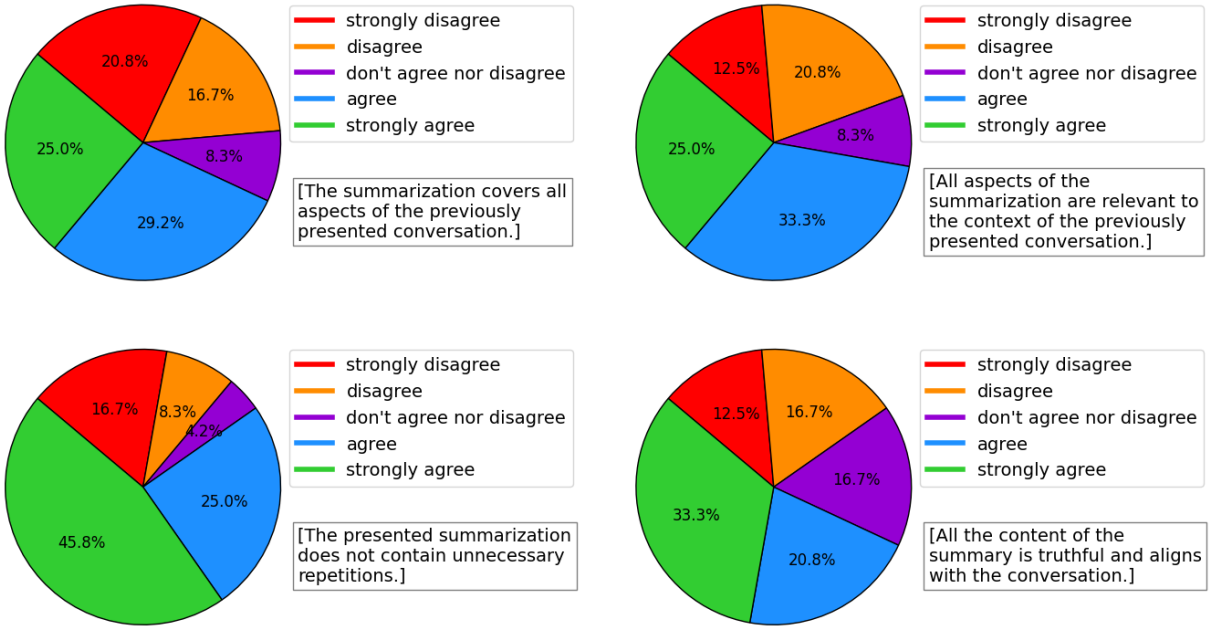


Observe que, em todas as perspectivas, as respostas positivas —- "concordo totalmente" e "concordo" combinadas —- representam mais de 50% do total de respostas. Mais especificamente, as porcentagens são de 62,5%, 58,3%, 54,2% e 70,9%, respectivamente, para cobertura, relevância, redundância e veracidade em relação aos resumos produzidos pelo LLaMA 3. Além disso, as respostas "discordo totalmente" são minoritárias em todas as perspectivas e está ausente em cobertura. Ou seja, os participantes consideraram os resumos do LLaMA 3 predominantemente satisfatórios em praticamente todas as perspectivas.

Para a Qwen 2, a Figura 7 também mostra uma predominância de respostas positivas, que, assim como para o LLaMA 3, representam mais de 50% do total. Ou seja, as porcentagens de respostas positivas são de 54,2%, 58,3%, 70,8% e 54,1%, para cobertura, relevância, redundância e veracidade, respectivamente, sugerindo também uma boa capacidade de sumarização para este LLM. No entanto, as respostas "discordo totalmente" para Qwen

2 são significativamente mais frequentes do que para LLaMA 3, ultrapassando 10% em todas as perspectivas.

Figura 7 – Distribuição das pontuações para cada perspectiva (24 avaliações por perspectiva), avaliando resumos gerados pelo Qwen 2. Respostas positivas — "concordo" e "concordo totalmente" — representam mais de 50% do total de respostas.



Análise de métricas estatísticas

A Tabela 5 mostra as taxas médias para cada perspectiva organizada por LLM. Todos os valores estão acima de 3, indicando uma tendência a resultados positivos e sugerindo uma boa eficácia de ambos os modelos (média). Para cobertura e relevância, LLaMA 3 obtém valores médios ligeiramente superiores aos obtidos por Qwen 2. Por outro lado, Qwen 2 se destaca pela (falta de) redundância em relação ao LLaMA 3. O oposto é verdadeiro para veracidade, indicando que, em média, LLaMA 3 produz resumos mais verdadeiros que Qwen 2.

Tabela 5 – Valores médios para cada perspectiva por LLM. Desvio-padrão entre parênteses.

Perspectiva	LLaMA 3	Qwen 2
Cobertura	3,29 (1,16)	3,21 (1,53)
Relevância	3,50 (0,98)	3,38 (1,41)
Redundância	3,42 (1,50)	3,75 (1,54)
Veracidade	3,83 (1,34)	3,46 (1,44)

Observa-se também que, para todas as perspectivas, os desvios-padrão obtidos usando o LLaMA 3 são menores, indicando menor dispersão dos dados. Isso sugere menor variação

nas respostas dos participantes, apontando para um maior nível de consenso, crucial para a seleção de um modelo. Isso será aprofundado a seguir.

Por fim, não foi observada nenhuma correlação entre os tamanhos dos diálogos e resumos com as avaliações qualitativas conduzidas em neste estudo. Isso sugere que o tamanho do resumo, dentro da faixa observada, não influencia significativamente as perspectivas avaliadas.

Kappa-Cohen (concordância entre os participantes)

Visto que o coeficiente Kappa-Cohen só é significativo quando calculado em avaliações realizadas sobre o mesmo objeto (resumo), foram selecionados quatro resumos gerados pelo LLaMA 3, representados na Tabela 6 como s_1, s_2, s_3 e s_4, e quatro resumos gerados pelo Qwen 2, representados como s_5, s_6, s_7 e s_8, totalizando oito resumos. Como três participantes avaliaram cada resumo, há três valores de Kappa-Cohen para cada resumo e, na Tabela 6, é mostrado o valor médio desses três valores para cada resumo.

Tabela 6 – Valores médios do coeficiente de Kappa-Cohen para cada LLM por resumo.

Resumo	LLaMA 3
s_1	0,116
s_2	0,036
s_3	0,283
s_4	0,087

Summary	Qwen 2
s_8	0,313
s_7	-0,022
s_6	0,056
s_5	-0,132

Os valores de Kappa-Cohen variam de -1 a 1, com valores menores que 0 indicando que a concordância é pior do que o esperado ao acaso. Observa-se tal situação nas avaliações dos resumos s_5 e s_7 gerados por Qwen 2. Valores entre 0,0 e 0,2 indicam concordância fraca, o que observa-se em vários resultados. Valores entre 0,2 e 0,4 indicam concordância razoável, como ocorre em s_3 e s_8 gerados por LLaMA 3 e Qwen 2, respectivamente.

De modo geral, ao analisar os valores obtidos pela aplicação do coeficiente Kappa-Cohen, nota-se que os valores de concordância foram geralmente baixos, com os resultados do LLaMA 3 sendo mais consistentes do que os do Qwen 2. No entanto, é importante ressaltar que, como essa métrica considera apenas a concordância total (mesma resposta), ela é muito sensível ao fato de haver cinco respostas possíveis. Em outras palavras, não há consideração do *grau de gradação* das respostas, sendo a discordância entre "concordo" e "concordo totalmente" considerada a mesma entre "concordo totalmente" e "discordo totalmente", por exemplo.

Análise qualitativa externa

Com o objetivo de obter uma visão de um profissional recém-integrado à equipe de saúde, foi solicitado ao participante externo — neste caso, um psicólogo que não fazia

parte da equipe de saúde da Ana Health — que avaliasse a utilidade dos resumos em auxiliar os colaboradores recém admitidos à equipe a compreender o contexto, quando os pacientes enviam novas mensagens. Assim, por meio de dois diálogos enviados como entrada tanto ao Qwen 2 quanto ao LLaMA 3, foram gerados dois resumos para cada diálogo. O participante avaliou os resumos quanto à clareza, coesão e utilidade dos resumos, sem saber quais LLMs foram utilizados.

Ao avaliar os resumos gerados pelo primeiro LLM (Qwen 2), o profissional considerou o primeiro resumo satisfatório, pois refletia com precisão os dados discutidos na conversa entre o paciente e a equipe de saúde. No entanto, o participante considerou o segundo resumo, também produzido por Qwen 2, insatisfatório, pois carecia de dados importantes para a compreensão completa do caso, que poderia impossibilitar um atendimento atencioso e eficiente.

Além disso, segundo o participante, o segundo resumo carecia do aspecto humano e não atendia aos requisitos para a compreensão do caso, visto que não fornecia dados cruciais, como o fato do paciente estar se recuperando da COVID, por exemplo. Além disso, o resumo continha detalhes sobre como o paciente havia expressado gratidão pelo atendimento, o que foi considerado sem importância para a compreensão da condição do paciente.

Em relação aos resumos gerados com o segundo LLM (LLaMA 3), o participante considerou que ambos os resumos mantiveram um certo padrão de qualidade e abordaram o conteúdo das conversas de forma eficiente. Nesse sentido, as quatro perspectivas solicitadas anteriormente foram atendidos, pois todos os aspectos da conversa foram bem resumidos, tiveram a relevância necessária para a compreensão do caso, foram objetivos e sem repetições desnecessárias, e não apresentaram conteúdo falso ou duvidoso.

Segundo o participante, resumos com avaliações positivas sob as perspectivas avaliadas podem auxiliar um novo membro da equipe a compreender o contexto de novas mensagens e, assim, formular respostas contextualizadas e personalizadas ao paciente com agilidade. Ele também informa que a capacidade de compreender o contexto de cada indivíduo, assim como a gestão do processo de acolhimento, é inerentemente pessoal e intransferível. “Cada pessoa é única e requer uma abordagem personalizada para o seu cuidado. O acolhimento personalizado, adaptado às necessidades e características específicas de cada indivíduo, é essencial para criar uma experiência verdadeiramente humanizada e eficaz”. De fato, o tempo costuma ser um fator crítico no processo de integração. O acesso mais rápido e preciso às informações melhora a eficiência desse processo (Alotaibi; Federico, 2017).

4 Discussão

Trabalhos anteriores sobre sumarização de dados clínicos não se aprofundaram na qualidade dos dados de entrada, que, neste estudo, como mostrado, podem ser precários e

informais. Esta análise em si é uma contribuição deste trabalho, pois impõe um desafio aos geradores de resumos baseados em LLMs. Além disso, há o fato dos diálogos serem escritos em português. De fato, embora seja uma das línguas mais faladas globalmente, o português tem sido historicamente sub-representado no treinamento de muitos grandes modelos de linguagens (LLMs) em comparação a idiomas como inglês, espanhol ou chinês. Isso pode ser atribuído a vários fatores, incluindo a disponibilidade de dados de treinamento de alta qualidade, o foco dos esforços de pesquisa e desenvolvimento, e o uso predominante do inglês na indústria de tecnologia (Bender et al., 2021).

Qualidade das mensagens

Em relação à dimensão tamanho, observa-se que a maioria das mensagens foi considerada curta (56,12%) e cerca de 23,58% foram consideradas longas. Textos curtos podem carecer de informações ou conter informações incompletas, enquanto textos longos podem apresentar redundâncias, o que pode levar os LLMs a não se concentrarem nas partes principais de um texto. Ambos os aspectos podem afetar a qualidade do resumo.

Os valores médios dos níveis Flesch-Kincaid da equipe de saúde e dos pacientes sugerem que um nível de escolaridade mais alto pode ser necessário para compreender as mensagens da equipe de saúde, em comparação com as mensagens dos pacientes. Em relação aos pacientes, observa-se que suas mensagens possuem uma faixa de valores mais ampla para os níveis de Flesch-Kincaid, do que as da equipe de saúde. Embora isso possa indicar uma potencial diversidade de pacientes em termos de nível de escolaridade, vários outros fatores também podem influenciar esses resultados, como o nível de detalhamento necessário para falar sobre determinado assunto e o uso de áudio para explicar assuntos mais complexos, entre outros.

Em relação à corretude, a maioria das palavras nas mensagens enviadas pela equipe de saúde é encontrada em um dicionário padrão, enquanto essa tendência não é encontrada em mensagens dos pacientes. De fato, a proporção de palavras "corretas" em mensagens de pacientes é muito baixa, em torno de 44% em média. No entanto, esse baixo nível de correção pode ser considerado e analisado neste contexto com mais cuidado, pois o uso da linguagem dos usuários do WhatsApp tem seus próprios termos, como "vc" (abreviação de "você"), "tbm" (abreviação de "também"), "pq" (abreviação de "porque"), que, da perspectiva dos usuários, são comuns e vistos como corretos. Em qualquer caso, a alta ocorrência de palavras incorretas ou não padronizadas pode levar os LLMs a ignorá-las ou interpretá-las mal para gerar os resumos, causando omissões de partes importantes do texto original ou geração de informações erradas.

Considerando as dimensões avaliadas, tamanho, legibilidade e corretude, há indícios de baixa qualidade nas mensagens dos pacientes. Essa (baixa) qualidade das mensagens em nosso conjunto de dados pode representar desafios significativos para a tarefa de sumarização usando LLMs.

Qualidade da tarefa de sumarização

Os resumos foram avaliados sob quatro perspectivas: cobertura, relevância, redundância e veracidade, por meio de uma pesquisa online. Ambos os LLMs obtiveram bom desempenho em todas as quatro perspectivas. Mais de 50% das respostas foram "concordo totalmente" e "concordo" em todas as perspectivas e em ambos os LLMs. Isso demonstra a capacidade dos LLMs de gerar resumos satisfatórios, apesar da qualidade do texto original. As poucas respostas "discordo totalmente" obtidas em meste trabalho (menos de 10%) concentram-se nos resumos gerados pelo Qwen 2, estando completamente ausente na perpectiva cobertura dos resumos gerados pelo LLaMA 3.

Comparando os resultados do LLaMA 3 com os do Qwen 2, considerando as respostas "concordo totalmente" e "concordo", nota-se que, em termos das perspectivas cobertura e veracidade, os resumos do LLaMA 3 são ligeiramente superiores aos produzidos pelo Qwen 2. Ambos os LLMs empatam sob a perspectiva de relevância e os resumos do Qwen 2 são ligeiramente superiores em relação à (ausência de) redundância. No geral, considerando todos os critérios de avaliação, o desempenho da sumarização do LLaMA 3, em nossa avaliação, é ligeiramente superior à do Qwen 2.

Em uma análise qualitativa externa, vista da perspectiva de um profissional recém-integrado à equipe de saúde, o profissional observou que os resumos produzidos por LLaMA 3 são objetivos, livres de repetições irrelevantes e não contêm conteúdo falso ou duvidoso. Por outro lado, os resumos produzidos por Qwen 2 não atendem ao mesmo padrão. Enquanto um resumo reflete corretamente o conteúdo do diálogo, o outro carece de dados importantes presentes no diálogo.

Avaliações positivas das perspectivas avaliadas podem significar que, os resumos produzidos podem efetivamente ajudar membros da equipe de saúde a rapidamente compreender o contexto de uma nova mensagem, sem a necessidade de rever todo o diálogo anterior com o paciente. Isso também permite que membros da equipe de saúde respondam a novas mensagens de maneira eficiente e contextualmente adequada. Essa agilidade no fornecimento de informações claras e precisas pode ajudar a reduzir a ansiedade dos pacientes que aguardam uma resposta, tornando o processo de interação entre pacientes e membros da equipe de saúde mais eficiente.

Resumidamente, apesar das mensagens serem escritas em português, alguns indícios de qualidade limitada das mensagens e alta variabilidade no nível de escolaridade necessário para entender as mensagens, os resumos produzidos por ambos os LLMs avaliados são satisfatórios para a tarefa alvo, com LLaMA 3 mostrando ligeira superioridade.

Limitações

Este trabalho apresenta algumas limitações. Primeiro, foram gerados resumos utilizando mensagens trocadas exclusivamente entre pacientes e a equipe de saúde da Ana Health.

O objetivo deste trabalho é apenas avaliar a capacidade dos LLMs em gerar resumos. Se fosse pretendido criar resumos contendo informações completas do paciente, como histórico de saúde, preferências e detalhes demográficos, serão necessárias fontes de dados adicionais. Segundo, a avaliação dos resumos foi realizada com membros da equipe de saúde da Ana Health. Talvez pessoas de outras áreas ou mesmo os pacientes possam ter opiniões diferentes. Terceiro, os resumos foram avaliados utilizando apenas a escala Likert. Se fosse pretendido entender em profundidade o motivo de cada valor registrado pelo participante, seriam necessários outros tipos de respostas, como comentários sobre os resumos avaliados. Quarto, a avaliação utilizou mensagens dos diálogos de oito pacientes como fonte. Uma análise mais abrangente da capacidade de sumarização dos LLMs pode exigir um número maior de diálogos. Quinto, este trabalho não avaliou o uso de resumos em outras tarefas, como seu impacto na resposta às mensagens dos pacientes.

Direções futuras

Há diversas possibilidades para aprimorar o desempenho e a aplicabilidade do sistema de sumarização em trabalhos futuros. Alguns esforços incluem a incorporação de prontuários médicos eletrônicos para enriquecer os dados de entrada, a integração do processamento de áudio para capturar e utilizar a comunicação verbal, e a adoção de estratégias que permitam a análise e a sumarização separadas de segmentos distintos dos dados antes de gerar de um resumo final abrangente.

Além disso, a análise qualitativa da capacidade dos LLMs em resumir diálogos foi limitada deliberadamente à quantidade de dados de entrada, para garantir que permanecessem dentro da janela de contexto dos modelos. Nesta fase, isso foi considerado uma restrição razoável, pois o objetivo principal era avaliar a eficácia com que os modelos poderiam capturar a essência das interações mais recentes entre pacientes e a equipe de saúde -— partindo do pressuposto de que mensagens recentes têm maior probabilidade de refletir o estado de saúde atual do paciente. Como um trabalho futuro, pretende-se investigar métodos para abordar as limitações impostas pelo tamanho da janela de contexto de grandes modelos de linguagem, que continua sendo um fator crítico, podendo influenciar a qualidade da sumarização (Kotkar et al., 2024; Zhang; Yu; Zhang, 2025).

Neste trabalho, um psicólogo externo avaliou resumos de acordo com o mesmo protocolo aplicado a todos os participantes. No futuro, também pretende-se incluir mais avaliadores externos para aprimorar a avaliação.

5 Conclusões

Neste trabalho, foram avaliadas as capacidades de dois grandes modelos de linguagem, LLaMA 3 e Qwen 2, em resumir diálogos entre pacientes e equipe de saúde. Esses diálogos, escritos em português, foram conduzidos via WhatsApp. Idealmente, os resumos gerados

devem destacar informações importantes sobre cada paciente, permitindo que a equipe de saúde responda a novas mensagens de pacientes de forma rápida e personalizada.

Diferentemente de trabalhos anteriores, que aplicaram sumarização em dados clínicos de alta qualidade (por exemplo, prontuários eletrônicos de saúde), este trabalho começa contribuindo com a avaliação da qualidade dos diálogos a serem sumarizados e foram encontradas evidências de baixa qualidade dos dados, especialmente devido à informalidade. O objetivo principal deste trabalho é avaliar rigorosamente se os LLMs estado-da-arte — especificamente modelos de código aberto — são capazes de produzir resumos clinicamente úteis a partir de diálogos reais, informais e ruidosos, entre pacientes e profissionais de saúde em português, um idioma sub-representado na pesquisa atual sobre processamento de linguagem natural aplicada à área de saúde.

A principal contribuição deste trabalho está na aplicação e avaliação sistemática desses modelos em um cenário realista de comunicação em saúde, que apresenta vários desafios técnicos e práticos: os diálogos são desestruturados, coloquiais e frequentemente fragmentados, refletindo padrões reais de comunicação em plataformas de saúde digital. O cenário envolve diálogos assíncronos, que evoluem cronológicamente, exigindo que os modelos mantenham coerência contextual sobre várias mensagens. Este domínio envolve conteúdo sensível relacionado à saúde, sendo que erros nos resumos podem ter implicações não triviais.

Endereçando as preocupações relacionadas a robustez e segurança dos resumos gerados, a avaliação neste trabalho foi estruturada em torno de quatro critérios principais: cobertura, relevância, redundância e veracidade. Essas dimensões foram escolhidas precisamente para avaliar potenciais erros de sumarização que poderiam impactar a interpretação clínica. Todas as avaliações foram conduzidas manualmente por humanos familiarizados com a área da saúde e nativos em português. Os resultados sugerem que, embora ambos os modelos tenham um desempenho razoavelmente bom, existem diferenças notáveis em sua capacidade de gerar resumos verdadeiros e relevantes — questões críticas em qualquer aplicação relacionada à saúde.

Resumidamente, a novidade deste trabalho reside em: (1) a aplicação e avaliação comparativa de LLMs em um domínio de poucos recursos e alto risco, utilizando diálogos clínicos em português do mundo real; (2) o desenvolvimento de uma estrutura de avaliação multifacetada, especificamente adaptada para identificar erros que possam comprometer tomadas de decisões clínicas; e (3) o estabelecimento de uma base para validação clínica futura, que abordará diretamente as preocupações de utilidade e segurança do paciente.

Concluindo, mesmo processando textos em português de baixa qualidade, os LLMs avaliados conseguiram produzir resumos que podem melhorar significativamente o atendimento ao paciente. Esses resumos podem auxiliar membros da equipe de saúde a compreender o contexto de novas mensagens enviadas pelos pacientes, permitindo responder de forma rápida e personalizada, melhorando, assim, a comunicação com os pacientes.

De uma perspectiva mais ampla, este trabalho destaca o papel crucial dos LLMs no enfrentamento das disparidades na área da saúde. Em países como o Brasil, onde uma parcela significativa da população não tem acesso a cuidados básicos de saúde e há escassez de profissionais de saúde, os serviços digitais de saúde não são apenas essenciais, mas representam uma solução crucial. Estratégias, como as discutidas neste trabalho, que podem processar efetivamente grandes volumes de dados em línguas sub-representadas, têm o potencial de causar um impacto social significativo, ajudando a aliviar questões urgentes.

Agradecimentos

Gostaríamos de agradecer à Ana Health, que faz parte desta colaboração, por apoiar este projeto. Agradecemos também à equipe da Ana Health. Este estudo foi apoiado também pelo Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (SEBRAE) e pela Empresa Brasileira de Pesquisa e Inovação Industrial (EMBRAPII). Também foi parcialmente financiado pela CAPES, CNPq, FAPEMIG e FINEP.

Referências

ALOTAIBI, Y. K.; FEDERICO, F. The impact of health information technology on patient safety. *Saudi medical journal*, Saudi Medical Journal, v. 38, n. 12, p. 1173, 2017. 20

BANERJEE, S.; LAVIE, A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. [S.l.: s.n.], 2005. p. 65–72. 9

BENDER, E. M. et al. On the dangers of stochastic parrots: Can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. [s.n.], 2021. (FAccT '21), p. 610–623. ISBN 9781450383097. Disponível em: <https://doi.org/10.1145/3442188.3445922>. 21

DALIP, D. H. et al. Automatic quality assessment of content created collaboratively by web communities: a case study of wikipedia. In: *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital libraries*. [S.l.: s.n.], 2009. p. 295–304. 5, 6

DAVE, T.; ATHALURI, S. A.; SINGH, S. Chatgpt in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in artificial intelligence*, Frontiers Media SA, v. 6, p. 1169595, 2023. 3

DILLMAN, D. A.; SMYTH, J. D.; CHRISTIAN, L. M. Internet, phone, mail, and mixed-mode surveys: The tailored design method. *John Wiley and Sons*, 2014. 10

DODDINGTON, G. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: *Proceedings of the second international conference on Human Language Technology Research*. [S.l.: s.n.], 2002. p. 138–145. 9

EL-KASSAS, W. S. et al. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, Elsevier, v. 165, p. 113679, 2021. 3, 9

FABBRI, A. R. et al. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 9, p. 391–409, 2021. 9

FERREIRA, A. A. et al. A comprehensive qualitative analysis of patient dialogue summarization using large language models applied to noisy, informal, non-english real-world data. *Scientific Reports*, Nature Publishing Group UK London, v. 15, n. 1, p. 31660, 2025. 1

GAO, M. et al. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*, 2023. 9

HONE, T. et al. Association between expansion of primary healthcare and racial inequalities in mortality amenable to primary care in brazil: a national longitudinal analysis. *PLoS medicine*, Public Library of Science San Francisco, CA USA, v. 14, n. 5, p. e1002306, 2017. 2

JAIN, S. et al. Multi-dimensional evaluation of text summarization with in-context learning. In: ROGERS, A.; BOYD-GRABER, J.; OKAZAKI, N. (Ed.). *Findings of the Association for Computational Linguistics: ACL 2023*. [s.n.], 2023. p. 8487–8495. Disponível em: <https://aclanthology.org/2023.findings-acl.537/>. 9

KESZTHELYI, D. et al. Patient information summarization in clinical settings: Scoping review. *JMIR Medical Informatics*, v. 11, p. e44639, Nov 2023. ISSN 2291-9694. Disponível em: <https://medinform.jmir.org/2023/1/e44639>. 3, 4

KINCAID, J. P. et al. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. [S.l.], 1975. 6

KOTKAR, A. D. et al. Comparative analysis of transformer-based large language models (llms) for text summarization. In: *2024 1st International Conference on Advanced Computing and Emerging Technologies (ACET)*. [S.l.: s.n.], 2024. p. 1–7. 23

LASKAR, M. T. R. et al. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In: AL-ONAIZAN, Y.; BANSAL, M.; CHEN, Y.-N. (Ed.). *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. [s.n.], 2024. p. 13785–13816. Disponível em: <https://aclanthology.org/2024.emnlp-main.764>. 3

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. [s.n.], 2004. p. 74–81. Disponível em: <https://aclanthology.org/W04-1013/>. 9

LIU, S. et al. Leveraging large language models for generating responses to patient messages—a subjective analysis. *Journal of the American Medical Informatics Association*, Oxford University Press, v. 31, n. 6, p. 1367–1379, 2024. 2

MANI, I.; MAYBURY, M. T. *Advances in Automatic Text Summarization*. Cambridge, MA, USA: MIT Press, 1999. ISBN 0262133598. 3

MARTINS, T. B. F. et al. *Readability formulas applied to textbooks in brazilian portuguese*. [S.l.], 1996. 6

MINAEE, S. et al. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024. 3

MORENO, G. C. d. L. et al. Alt: A software for readability analysis of portuguese-language texts. *arXiv preprint arXiv:2210.00553*, 2022. 6

NIELSEN, J. *Why you only need to test with 5 users. In Nielsen Norman Group*. 2000. Https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/. 10

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. [s.n.], 2002. p. 311–318. Disponível em: <https://doi.org/10.3115/1073083.1073135>. 9

REVILLA, M.; OCHOA, C. Ideal and maximum length for a web survey. *International Journal of Market Research*, SAGE Publications Sage UK: London, England, v. 59, n. 5, p. 557–565, 2017. 10

WANG, L. et al. Applications and concerns of chatgpt and other conversational large language models in health care: Systematic review. *Journal of Medical Internet Research*, JMIR Publications Toronto, Canada, v. 26, p. e22769, 2024. 3

WU, N. et al. Large language models are diverse role-players for summarization evaluation. In: *CCF International Conference on Natural Language Processing and Chinese Computing*. [S.l.: s.n.], 2023. p. 695–707. 9

YANG, R. et al. Large language models in health care: Development, applications, and challenges. *Health Care Science*, Wiley Online Library, v. 2, n. 4, p. 255–263, 2023. 3

YUAN, W.; NEUBIG, G.; LIU, P. BARTSCORE: evaluating generated text as text generation. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. [S.l.: s.n.], 2021. (NIPS '21). ISBN 9781713845393. 9

ZHANG, H.; YU, P. S.; ZHANG, J. A systematic survey of text summarization: From statistical methods to large language models. *ACM Computing Surveys*, abr. 2025. ISSN 0360-0300. Just Accepted. Disponível em: <https://doi.org/10.1145/3731445>. 3, 23

ZHANG, T. et al. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. Disponível em: <https://doi.org/10.48550/arXiv.1904.09675>. 9