UNIVERSIDADE FEDERAL DE JUIZ DE FORA
FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

João Pedro Junqueira Schettino

Grandes modelos de linguagem aplicados a intervenções em saúde mental

Juiz de Fora
2024

João Pedro Junqueira Schettino

Grandes modelos de linguagem aplicados a intervenções em saúde mental

Dissertação apresentada ao Programa de Pós Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Orientador: Dr. Heder Soares Bernardino

Coorientador: Dr. Jairo Francisco de Souza

Juiz de Fora

2024

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Schettino, Joao.
Grandes modelos de linguagem aplicados a intervenções em saúde mental / Joao Schettino. -- 2024.
110 f.

Orientadora: Heder Soares Bernardino
Coorientadora: Jairo Francisco de Souza
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Faculdade de Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2024.

1. Grandes Modelos de Linguagem. 2. Intervenção em saúde mental. I. Soares Bernardino, Heder, orient. II. Francisco de Souza, Jairo, coorient. III. Título.

João Pedro Junqueira Schettino

Grandes Modelos de Linguagem Aplicados a Intervenções de Saúde Mental

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 19 de dezembro de 2024.

BANCA EXAMINADORA

Prof. Dr. Heder Soares Bernardino - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Jairo Francisco de Souza - Coorientador
Universidade Federal de Juiz de Fora

Prof. Dr. Leonardo Goliatt da Fonseca
Universidade Federal de Juiz de Fora

Prof. Dr. Carlos Eduardo Raymundo
Universidade do Estado do Rio de Janeiro

Juiz de Fora, 17/12/2024.

Documento assinado eletronicamente por **Heder Soares Bernardino**, **Professor(a)**, em 19/12/2024, às 16:48, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Jairo Francisco de Souza**, **Professor(a)**, em 19/12/2024, às 17:27, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca**, **Professor(a)**, em 20/12/2024, às 10:26, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Carlos Eduardo Raymundo**, **Usuário Externo**, em 14/01/2025, às 08:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2162816** e o código CRC **D58AC0CA**.

AGRADECIMENTOS

Gostaria de expressar minha mais profunda gratidão à minha família, cuja influência foi fundamental na minha formação pessoal e no meu desenvolvimento enquanto indivíduo. Aos amigos, meu sincero agradecimento pelo apoio constante e pela indispensável validação ao longo desta jornada.

Sou imensamente grato aos meus professores orientadores, Heder Soares Bernardino e Jairo Francisco de Souza, por sua paciência, perspicácia, valiosas reflexões e por guiarem meu trabalho com dedicação e profissionalismo. Não poderia deixar de reconhecer a contribuição de Leonardo Fernandes Martins, cuja importância foi singular para este trabalho, mesmo que, por questões formais, não tenha sido oficialmente reconhecido como coorientador.

Registro ainda minha gratidão ao grupo Álcool e Saúde, que tanto contribuiu para este percurso, à Universidade Federal de Juiz de Fora e ao Programa de Pós-Graduação em Modelagem Computacional, pela infraestrutura de excelência e pela formação acadêmica que proporcionaram. Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, pelo apoio financeiro essencial ao desenvolvimento desta pesquisa, ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), cujos aportes foram fundamentais para a realização deste trabalho.

“O correr da vida embrulha tudo. A vida é assim: esquenta e esfria, aperta e afrouxa, sossega e depois desinquieta. O que ela quer da gente é coragem.” – João Guimarães Rosa, Grande Sertão: Veredas

RESUMO

A crescente preocupação com a necessidade de intervenções eficazes em saúde mental é evidente, especialmente devido à escassez de profissionais qualificados para lidar com a demanda existente. A área da saúde mental muitas vezes se encontra em uma situação onde a procura por serviços supera significativamente a oferta, resultando em muitos indivíduos sem acesso ao suporte necessário. Diante desse cenário, é crucial explorar alternativas viáveis para auxiliar na prestação de cuidados. Dentre as diversas demandas presentes na área da saúde mental no Brasil, o abuso de álcool emerge como uma das grandes mazelas enfrentadas na saúde pública. Uma possível solução para esta problemática surge com o uso de grandes modelos de linguagem para compreender e gerar texto de forma contextualmente relevante e precisa. Esses modelos podem desempenhar um papel significativo na ampliação do acesso aos serviços de saúde mental. Esse tipo de abordagem não só pode ajudar a aliviar a pressão sobre os profissionais de saúde mental, mas também a alcançar uma maior parcela da população que pode não ter acesso ao cuidado especializado. Esta dissertação explora a proposta de uma ferramenta de auxílio à intervenção para a saúde mental, no contexto do uso de álcool, fornecendo auxílio ao consultor para seguir as melhores práticas. Para entender melhor as ferramentas disponíveis e buscar os caminhos mais apropriados, foi realizada uma revisão sistemática. Essa revisão teve como objetivo mapear as características dos *chatbots* voltados para a saúde mental, incluindo aspectos de usabilidade, tecnologias e metodologias utilizadas, métodos de avaliação e estágio de desenvolvimento dos estudos. A partir dessas descobertas, propõe-se explorar a aplicação dos grandes modelos de linguagem LLAMA e GEMMA no contexto da intervenção sobre Álcool e Saúde. Em um primeiro momento, é realizada uma modelagem com dados disponibilizados de forma pública, o que valida a nossa proposta de solução entregando resultados satisfatórios, mesmo com a utilização de dados traduzidos do inglês para fazer o treinamento do modelo. Em um segundo momento, o processo de treinamento do modelo, bem como das adaptações realizadas para comportar nossos objetivos é demonstrado. Utilizamos dados de sessões de psicoterapia reais. A avaliação dos modelos foi realizada por avaliadores especialistas no domínio do problema. Este trabalho contribui para a compreensão do estado atual dos *chatbots* aplicados à saúde mental. Além disso, a comparação e o refinamento dos modelos LLAMA e GEMMA promovem o entendimento sobre o desempenho de modelos de linguagem em contextos específicos de saúde mental

Grandes Modelos de Linguagem: Intervenção em saúde mental; Consumo excessivo de Álcool; Chatbots.

ABSTRACT

Growing concern about the need for effective mental health interventions is evident, especially given the shortage of qualified professionals to deal with the existing demand. The mental health field often finds itself in a situation where demand for services significantly outstrips supply, resulting in many individuals not having access to the necessary support. Faced with this scenario, it is crucial to explore viable alternatives to help provide care. Among the various demands present in the area of mental health in Brazil, alcohol abuse emerges as one of the major problems faced as a public health issue. A possible solution to this problem arises with the use of large language models to understand and generate text in a contextually relevant and accurate way. These models can play a significant role in expanding access to mental health services. This approach can not only help relieve pressure on mental health professionals, but also reach a larger portion of the population who may not have easy access to therapists or specialised clinics. This dissertation explores the proposal of an intervention aid tool for mental health, especially in the context of alcohol use, providing help to the counsellor to follow best practices. In order to better understand the tools and look for the most appropriate ways forward, a systematic review was carried out. This review aimed to map the characteristics of *chatbots* aimed at mental health, including aspects of usability, technologies and methodologies used, evaluation methods and the stage of development of studies. Based on these findings, the dissertation proposes to explore the application of the major language models LLAMA and GEMMA in the context of brief intervention on Alcohol and Health. Firstly, modelling is carried out using publicly available data. This validates our proposed solution by delivering satisfactory results, even when using data translated from English to train the model. Secondly, the process of training the model and adapting it to meet our objectives is demonstrated. We used data from real therapy sessions. The models are evaluated by expert evaluators in the problem domain. This work contributes to understanding the current state of *chatbots* applied to mental health. In addition, the comparison and refinement of the LLAMA and GEMMA models promotes understanding of the performance of language models in specific mental health contexts

LLM: Mental Health Intervention; Alcohol abuse; Chatbots.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AC	Agentes Conversacionais
AI	Artificial Intelligence
BLEU	Bilingual Evaluation Understudy
ChatGPT	Chat Generative Pre-trained Transformer
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DP	Desvio Padrão
EM	Entrevista Motivacional
ER	Reflexão Explicativa
EX	Extensividade de Respostas
GPT	Generative Pre-trained Transformer
HEAL	Health Emotional Assistance Logic
IP	Proporção de Interações
JSON	JavaScript Object Notation
LLM	Large Language Model
LoRA	Low-Rank Adaptation
LSTM	Long Short-Term Memory
MAE	Avaliações Baseadas em Modelo e Métricas Automatizadas
ML	Machine Learning
MME	Avaliações de Métodos Mistos
NLP	Natural Language Processing
OMS	Organização Mundial da Saúde
OR	Pesquisa Observacional
PFET	Ajuste Fino Eficiente de Parâmetros
QA	Question Answering
RNN	Recurrent Neural Network
RoPE	Rotary Position *embedding*
Seq2Seq	Sequence to Sequence
SVM	Support Vector Machine
TCC	Terapia Cognitivo-Comportamental
TEPT	Transtorno do Estresse Pós-Traumático
UMAP	Uniform Manifold Approximation and Projection
UCE	Avaliações Focadas no Usuário
WDH	Whole Dialogue History

SUMÁRIO

1 INTRODUÇÃO

A saúde mental tem se consolidado como um dos maiores desafios da saúde pública contemporânea, enfrentando um panorama de alta demanda por serviços especializados e uma escassez crítica de profissionais qualificados para atender às necessidades da população. Essa discrepância tem se tornado cada vez mais evidente em diversas regiões do mundo, inclusive no Brasil, onde os recursos para saúde mental muitas vezes são insuficientes para atender às demandas crescentes (27). A Organização Mundial da Saúde (OMS) estima que os transtornos mentais sejam responsáveis por uma parcela significativa da carga global de doenças, afetando milhões de pessoas em todo o mundo. Neste cenário, o abuso de álcool se destaca como uma das principais causas de sofrimento psíquico e danos sociais, configurando-se como uma questão central de saúde pública (92).

O consumo problemático de álcool tem implicações diretas na saúde física e mental dos indivíduos, impactando também suas relações sociais, econômicas e familiares. Estudos apontam que o uso abusivo de álcool não apenas exacerba condições psiquiátricas preexistentes, mas também pode levar ao desenvolvimento de novos transtornos mentais, além de estar associado a comportamentos de risco, acidentes e violência (55). Essa realidade destaca a necessidade urgente de intervenções eficazes e acessíveis, capazes de lidar com a complexidade do problema e de alcançar populações vulneráveis de maneira ampla e equitativa.

Entretanto, as abordagens tradicionais em saúde mental, que dependem exclusivamente de interações presenciais entre pacientes e profissionais, apresentam limitações significativas em termos de alcance e sustentabilidade. Essa lacuna evidencia a necessidade de soluções inovadoras que utilizem tecnologia para democratizar o acesso ao cuidado em saúde mental. Neste contexto, a inteligência artificial (IA), e mais especificamente os grandes modelos de linguagem, surge como uma alternativa promissora. Esses modelos, que se destacam por sua capacidade de compreender e gerar textos de forma contextualmente relevante, oferecem novas possibilidades para apoiar intervenções em saúde mental (28).

Os grandes modelos de linguagem, como o LLAMA e o GEMMA (53, 73) , têm demonstrado um potencial significativo para atuar como ferramentas auxiliares em intervenções psicológicas. Suas capacidades avançadas de processamento de linguagem natural permitem interações fluidas e personalizadas, simulando a comunicação humana. Além disso, esses modelos apresentam características que os tornam especialmente adequados para aplicações sensíveis como as da saúde mental, incluindo a capacidade de adaptar respostas com base no contexto emocional e linguístico do usuário (53, 73).

A presente dissertação tem como objetivo principal investigar a aplicação desses grandes modelos de linguagem no contexto de intervenções para o uso problemático de álcool. O problema central abordado é a criação de uma ferramenta tecnológica capaz de

apoiar profissionais de saúde mental, facilitando o processo de intervenção e ampliando o acesso ao cuidado. Essa ferramenta, fundamentada em modelos como LLAMA e GEMMA, busca oferecer suporte contextualizado, empático e alinhado às melhores práticas em saúde mental. A escolha por esses modelos é motivada tanto por seus avanços tecnológicos quanto por seu potencial de impacto social, considerando as necessidades específicas do público-alvo e as limitações das abordagens atuais.

Para embasar a proposta, foi realizada uma revisão sistemática da literatura, com o objetivo de mapear o estado da arte em *chatbots* voltados para a saúde mental. Esta revisão permitiu identificar as características mais relevantes dessas ferramentas, incluindo aspectos de usabilidade, tecnologias empregadas, metodologias utilizadas para treinamento e avaliação, bem como as limitações presentes nos estudos existentes (108). A partir dessas descobertas, foi possível delinear diretrizes claras para o desenvolvimento da ferramenta proposta nesta dissertação.

Em sequência, o trabalho é dividido em duas etapas principais. Na primeira etapa, foi realizada uma modelagem inicial utilizando dados públicos de entrevistas motivacionais, traduzidos do inglês, para validar a proposta de solução e testar a capacidade dos modelos de gerar respostas contextualmente adequadas. Os resultados iniciais indicaram que, mesmo com dados limitados, os modelos foram capazes de oferecer respostas coerentes e úteis em cenários simulados. Na segunda etapa, o treinamento dos modelos foi aprimorado utilizando dados reais de sessões terapêuticas coletados com o consentimento dos participantes e após aprovação de um comitê de ética. Essas sessões seguem o protocolo MATCH (55), que é amplamente reconhecido por sua eficácia em intervenções breves no contexto do abuso de álcool.

A dissertação também explora técnicas de calibração e otimização dos modelos, incluindo métodos como escalonamento ROPE e LoRA (Low-Rank Adaptation), que são essenciais para ajustar os modelos às especificidades do domínio da saúde mental (20, 96). Além disso, são discutidos aspectos metodológicos relacionados ao tratamento dos dados, formulação do problema e avaliação dos resultados, utilizando transcrições de sessões terapêuticas reais e metodologias estatísticas rigorosas para análise (82).

Do ponto de vista acadêmico e social, este trabalho oferece contribuições relevantes para o campo da saúde mental e da inteligência artificial. Em primeiro lugar, a revisão sistemática realizada proporciona uma visão abrangente sobre o estado atual dos *chatbots* aplicados à saúde mental, identificando lacunas e oportunidades de pesquisa. Em segundo lugar, a aplicação prática dos modelos LLAMA e GEMMA no contexto de intervenções sobre o uso de álcool pode servir como um modelo para o desenvolvimento de tecnologias similares em outras áreas da saúde mental. Por fim, ao explorar e documentar metodologias de treinamento, calibração e avaliação de modelos de linguagem para aplicações sensíveis, este estudo estabelece um referencial para futuras pesquisas e desenvolvimento de soluções

tecnológicas.

Conclui-se que a combinação entre avanços tecnológicos e metodologias baseadas em evidências pode oferecer soluções significativas para os desafios da saúde mental no Brasil e no mundo. Ao propor uma abordagem para intervenções sobre o uso problemático de álcool, esta dissertação busca contribuir para a construção de um futuro mais inclusivo e acessível no cuidado em saúde mental.

2 REVISÃO DA LITERATURA

A aplicação de *chatbots* no contexto da saúde mental tem recebido crescente atenção nos últimos anos, impulsionada pelos avanços das tecnologias de inteligência artificial e pela demanda por soluções acessíveis e escaláveis no cuidado psicológico. Apesar do entusiasmo em torno dessas ferramentas, ainda existem lacunas importantes no entendimento de sua implementação, eficácia e impacto. Dado esse cenário, torna-se essencial uma revisão sistemática que examine as características subjacentes, as tecnologias empregadas e os modelos computacionais, bem como a robustez das evidências que sustentam a eficácia dos *chatbots* em saúde mental. Duas perguntas de pesquisa guiam este estudo:

- RQ1: Como variam os modelos de *chatbot* em sua implementação para diferentes casos de uso?
- RQ2: Quão robustas são as evidências que sustentam a eficácia desses *chatbots* e quais aspectos estão sendo avaliados?

A rápida evolução da área e o crescente reconhecimento da importância dos *chatbots* como ferramentas de saúde mental enfatizam a necessidade dessa revisão. Assim, nosso estudo busca não apenas catalogar e analisar as tendências atuais, mas também identificar lacunas na pesquisa existente e sugerir direções futuras para as aplicações desses desenvolvimentos clínicos e tecnológicos. Com isso, esperamos avançar na área ao garantir que as intervenções baseadas em *chatbots* sejam inovadoras e fundamentadas em evidências clínicas sólidas.

O panorama dos agentes conversacionais (ACs) para saúde mental abrange diversos domínios, incluindo ciência da computação e medicina, e vários estudos buscam explorar diferentes aspectos de seu uso, eficácia e desafios. Uma revisão sistemática realizada por (18) visa preencher a lacuna entre essas disciplinas, comparando direções de pesquisa e metodologias para aprimorar os métodos de pesquisa de ACs em saúde mental. Embora este estudo tenha semelhanças com o nosso quanto ao escopo da coleta de dados (estendendo-se até 2022), nossa revisão se distingue pelo foco em compreender a aplicação, avaliação e maturidade dos modelos de AC com pelo menos um componente textual. Diferente de (18), nossa pesquisa não visa comparar as áreas de enfoque dessas disciplinas, mas sim investigar como diferentes componentes dos modelos de AC são empregados e avaliados em diversos cenários.

Estudos que examinam o engajamento de usuários no mundo real com aplicativos de saúde mental, como em (8), destacam a diversidade de funcionalidades dos aplicativos e os comportamentos dos usuários. Seus achados indicam diferenças significativas na retenção e engajamento dos usuários em diferentes tipos de aplicativos, como aplicativos de suporte entre pares e de mindfulness, em comparação com rastreadores ou exercícios de respiração.

Embora sua análise forneça *insights* valiosos sobre os padrões de comportamento dos usuários, ela diverge de nossa revisão, que se interessa especificamente pelos componentes estruturais e funcionais dos modelos de AC, e não pelas métricas gerais de engajamento de aplicativos.

O impacto mais amplo dos aplicativos de saúde digital nos contextos de saúde mental é explorado em (6), que relata que pacientes que utilizam tais aplicativos frequentemente apresentam melhorias nas condições de saúde, com alta satisfação e usabilidade. No entanto, o foco em aplicativos de saúde digital em geral, ao invés de ACs com componentes específicos de modelo, diferencia este estudo do nosso. De forma semelhante, a revisão sistemática de (109) se concentra na atividade física, qualidade do sono e outros desfechos de saúde, fornecendo contexto para intervenções de saúde digital, mas fora do foco central dos modelos de AC na saúde mental.

Vários estudos enfatizam a experiência do usuário e os desafios associados aos *chatbots* de saúde mental. (41) fornecem uma visão geral dos *chatbots* de saúde mental disponíveis comercialmente, destacando tanto os benefícios das interações personalizadas quanto os riscos, como o apego do usuário e o manejo inadequado de respostas em crises. Essas percepções ressaltam a importância de avaliar a capacidade dos modelos de AC de atender efetivamente às necessidades dos usuários, uma preocupação também relevante para o exame da maturidade e dos métodos de avaliação dos ACs em nosso estudo.

Questões de privacidade associadas aos aplicativos de saúde mental são analisadas criticamente por (127), que explora os antecedentes dos medos dos pacientes em relação à privacidade e seu impacto na continuidade do uso dos serviços móveis de saúde mental. Suas descobertas destacam que questões relacionadas à privacidade afetam significativamente as atitudes e intenções de uso dos usuários, fator que indiretamente influencia a aceitação e a eficácia dos modelos de AC. No entanto, nossa revisão se concentra principalmente nos componentes e avaliação dos modelos de AC, ao invés da privacidade.

(60) categoriza os *chatbots* de saúde com base em seus papéis e limitações, identificando áreas-chave como suporte ao paciente, promoção de comportamento saudável e desafios éticos e técnicos. Essa categorização ampla ajuda a contextualizar as variadas funções dos ACs na saúde, mas não se aprofunda nos componentes específicos dos modelos de AC ou em sua avaliação, como nosso estudo faz. Finalmente, (7) aborda o *design* ético, responsável e confiável de *chatbots* de IA, apontando desafios como o viés algorítmico. Sua narrativa destaca o potencial e os riscos do suporte à saúde mental movido por IA, alinhando-se com nosso interesse mais amplo em avaliar a maturidade e a eficácia das implementações de AC.

Em resumo, enquanto a literatura existente fornece uma visão abrangente dos *chatbots* de saúde mental e intervenções digitais relacionadas, nossa revisão sistemática contribui de forma única ao focar nos componentes estruturais e funcionais dos ACs, como

são empregados, sua avaliação em diferentes contextos e a maturidade dos estudos nesse campo. Essa abordagem distinta oferece uma compreensão mais profunda das dinâmicas específicas dos modelos em agentes conversacionais voltados para a saúde mental.

2.1 CONTEXTUALIZAÇÃO

Chatbots, ou agentes conversacionais, são programas de software projetados para simular conversas com usuários humanos por meio de texto, fala ou interfaces visuais. Eles têm recebido significativa atenção no campo da saúde mental devido ao seu potencial de oferecer suporte escalável, acessível e econômico para indivíduos com preocupações relacionadas à saúde mental. Esta seção aborda os conceitos fundamentais, estruturas tecnológicas e aplicações práticas de *chatbots* na saúde mental.

2.1.1 Tipos de Modelos Conversacionais

Chatbots projetados para aplicações em saúde mental baseiam-se em diversos modelos computacionais, frequentemente incorporando múltiplos componentes adaptados a funcionalidades específicas. Esses componentes podem incluir modelos generativos, mecanismos de classificação de intenções, sistemas de reconhecimento de emoção ou sentimento, estruturas de recuperação de conhecimento e lógica baseada em regras para orientar conversas. Compreender a presença ou ausência desses componentes em diferentes modelos de *chatbots* pode revelar categorizações-chave que diferenciam suas capacidades.

O componente gerador em *chatbots* refere-se à capacidade do modelo de produzir novas respostas contextualmente relevantes, em vez de selecionar entre um conjunto de respostas predefinidas. Isso é alcançado por meio de modelos generativos, uma classe de modelos de aprendizado de máquina projetados para aprender padrões em dados de linguagem e gerar texto com base nessa distribuição aprendida. Arquiteturas comuns para modelos generativos incluem Sequence-to-Sequence (Seq2Seq) (11) e modelos baseados em Transformadores, como os Transformers Pré-Treinados Generativos (GPT) (129). Esses modelos criam respostas prevendo a próxima palavra ou token em uma sequência, com base no input fornecido pelo usuário, tornando-os altamente flexíveis para conversas abertas (119, 98).

Uma das principais vantagens dos modelos generativos é sua capacidade de lidar com uma ampla variedade de entradas e gerar respostas diversificadas. Por exemplo, modelos em larga escala como GPT-2 ou GPT-4, treinados em extensos conjuntos de dados, podem produzir diálogos semelhantes aos humanos em muitos contextos diferentes (119). Esses modelos não dependem de respostas predefinidas, permitindo que respondam a consultas novas ou inesperadas, oferecendo uma experiência de usuário mais dinâmica e personalizada (98).

No entanto, o uso de modelos generativos também apresenta certas limitações. Embora possam produzir respostas altamente fluentes e flexíveis, eles são propensos a gerar informações incorretas, sem sentido ou inconsistentes, particularmente em domínios especializados, como a saúde mental, onde precisão e confiabilidade são cruciais. Além disso, esses modelos frequentemente exigem recursos computacionais significativos para treinamento e implantação, o que pode não ser viável em todas as aplicações (50). Ao contrário de sistemas baseados em regras ou modelos de recuperação ancorados em bases de conhecimento estruturadas, os modelos generativos às vezes podem produzir respostas factualmente incorretas ou irrelevantes porque não estão explicitamente ancorados em uma estrutura de recuperação de conhecimento (30).

Em resumo, os modelos generativos oferecem vantagens consideráveis em flexibilidade e na capacidade de gerenciar diálogos abertos; no entanto, também apresentam desafios, particularmente em relação à confiabilidade e ao custo computacional. Sua aplicação em *chatbots* de saúde mental deve ser cuidadosamente gerenciada para garantir que as respostas geradas sejam apropriadas, precisas e apoiem o bem-estar do usuário.

A classificação de intenções é outro componente essencial, especialmente em *chatbots* baseados em regras ou específicos para tarefas. Esse mecanismo permite que o *chatbot* compreenda e categorize a entrada do usuário. A classificação de intenções ajuda a orientar a conversa ao determinar os objetivos ou necessidades subjacentes do usuário. Por exemplo, *chatbots* como os baseados no Dialogflow do Google (57) ou na estrutura RASA (93) utilizam esse componente para detectar intenções do usuário e fornecer respostas relevantes (63, 62). Em muitos casos, a classificação de intenções é suportada por modelos de aprendizado de máquina, como Máquinas de Vetores de Suporte (SVMs) (29), redes LSTM (Long Short-Term Memory) ou redes neurais profundas (DNNs) (21). Um *chatbot* que utiliza LSTM para classificação de intenções, como o descrito por Harilal et al. (42), pode melhorar a fluência conversacional ao manter o contexto ao longo do tempo.

O reconhecimento de emoção ou sentimento adiciona outra camada de sofisticação aos modelos de chatbots, permitindo-lhes responder de forma empática ao detectar o estado emocional do usuário. Este componente frequentemente emprega técnicas como análise de sentimentos ou redes neurais especializadas para detecção de emoções. *Chatbots* como o desenvolvido por (25) utilizam modelos de análise de sentimentos, como classificadores de Floresta Aleatória, para reconhecer emoções como tristeza, raiva ou alegria. Outros modelos podem integrar modelos probabilísticos ou ferramentas como o TextBlob para analisar texto em busca de pistas emocionais (21). Ao incorporar o reconhecimento de emoções, os *chatbots* podem personalizar suas respostas para fornecer suporte emocional, particularmente em cenários de saúde mental onde entender os sentimentos do usuário é crucial para uma interação eficaz.

A recuperação de conhecimento é outro componente crítico, especialmente para

chatbots que necessitam fornecer informações ou conselhos precisos. Essa funcionalidade permite que o *chatbot* recupere dados relevantes de uma base de conhecimento ou fontes externas. Por exemplo, o *chatbot* descrito por (77) utiliza uma rede de média profunda (DAN) para corresponder consultas de usuários a respostas relevantes de um conjunto de dados, garantindo que as respostas sejam fundamentadas em informações confiáveis. Em contraste, *chatbots* que utilizam grafos de conhecimento ou sistemas de recuperação de dados estruturados, como os implementados por (23), podem acessar conjuntos de dados complexos para fornecer respostas mais abrangentes e contextualmente apropriadas.

Por fim, a lógica de regras define como os *chatbots* gerenciam conversas por meio de árvores de decisão predefinidas ou máquinas de estados finitos. Isso é particularmente prevalente em modelos mais simples que dependem de lógica determinística em vez de aprendizado de máquina. Woebot, um conhecido *chatbot* de saúde mental descrito por (30), exemplifica essa abordagem, utilizando uma estrutura de árvore de decisão para navegar nas conversas com base na entrada do usuário. Sistemas baseados em regras como esses frequentemente garantem respostas consistentes, mas podem carecer da flexibilidade de modelos gerativos mais sofisticados. No entanto, são altamente eficazes para aplicações específicas onde o fluxo da conversa deve seguir diretrizes rigorosas, como a entrega de intervenções terapêuticas ou o seguimento de protocolos diagnósticos (103).

Em conclusão, os modelos computacionais de *chatbots* de saúde mental podem ser compreendidos ao examinar a presença ou ausência de componentes-chave, como mecanismos generativos, classificação de intenções, reconhecimento de emoções ou sentimentos, recuperação de conhecimento e lógica baseada em regras. Cada componente contribui para a funcionalidade geral do chatbot, permitindo que ele gere respostas, compreenda intenções do usuário, detecte estados emocionais, recupere informações relevantes ou siga caminhos conversacionais predefinidos. Ao adaptar a inclusão desses componentes ao propósito do chatbot, os desenvolvedores podem criar modelos que atendam às necessidades específicas dos usuários em contextos de saúde mental.

2.1.2 Os Papéis dos *Chatbots* na Saúde Mental

Chatbots na área de saúde mental frequentemente desempenham diferentes papéis, cada um adaptado para atender necessidades específicas dos usuários no contexto mais amplo da saúde mental. Um papel significativo dos *chatbots* está em fornecer *Intervenções Terapêuticas Autoguiadas*. Esses *chatbots* são projetados para oferecer técnicas terapêuticas, como Terapia Cognitivo-Comportamental (TCC) ou Entrevista Motivacional (EM), sem a necessidade de envolvimento direto humano. Por exemplo, o Woebot (30) entrega TCC por meio de interações breves e diárias para ajudar os usuários a gerenciar depressão e ansiedade. De forma semelhante, o Bonobot (88) utiliza técnicas de EM para orientar estudantes de pós-graduação na gestão do estresse, promovendo autorreflexão e encorajando mudanças

positivas. Essas intervenções são estruturadas e baseadas em evidências, oferecendo aos usuários uma experiência terapêutica semelhante à terapia humana, mas de forma mais acessível, frequentemente assíncrona. Outro chatbot, o SELMA (44), promove o autogerenciamento de dores crônicas por meio de técnicas de TCC, ajudando os usuários a mudar cognições disfuncionais e lidar mais eficazmente com sua condição.

Outro papel chave está em *Apoio Psico-Informativo*, onde *chatbots* fornecem educação em saúde mental, orientação ou informações médicas relevantes aos usuários. Esse suporte visa melhorar a compreensão dos usuários sobre suas condições e oferecer conselhos práticos. Por exemplo, o Vik (15) foi projetado para capacitar pacientes com doenças crônicas ao fornecer informações personalizadas para melhorar a adesão ao tratamento e a qualidade de vida. De forma semelhante, o Robo (77) ajuda pacientes viciados em opioides ao buscar informações relevantes em plataformas de mídia social como o Reddit, oferecendo respostas rápidas para dúvidas de suporte. Esses *chatbots* se concentram em entregar informações precisas e oportunas, ajudando os usuários a tomar decisões informadas sobre sua saúde mental e tratamento.

Uma categoria fundamental é *Triagem e Avaliação*, onde os *chatbots* são utilizados para detectar, triar ou avaliar condições de saúde mental. Esses *chatbots* normalmente usam ferramentas estruturadas, como escalas psicológicas validadas ou processamento de linguagem natural, para analisar as entradas dos usuários. Por exemplo, um *chatbot* desenvolvido por (121) monitora a saúde mental de mulheres perinatais, avaliando ansiedade, depressão e hipomania, ao mesmo tempo em que fornece recomendações médicas em tempo real baseadas no estado mental do usuário. De forma semelhante, um *chatbot* desenvolvido por (43) oferece uma avaliação inicial de saúde mental utilizando testes de personalidade e experimentos psicológicos, ajudando os usuários a autoavaliar suas condições de saúde mental. Essas avaliações ajudam a preencher a lacuna entre os usuários e os profissionais, fornecendo intervenções precoces e identificação oportuna de problemas de saúde mental.

Chatbots também desempenham um papel vital em *Apoio Emocional e Empatia*, oferecendo aos usuários diálogos empáticos, conforto emocional e processamento emocional. Esses *chatbots* têm como objetivo proporcionar uma presença reconfortante para usuários que enfrentam sofrimento emocional, frequentemente atuando como o primeiro ponto de contato emocional. Por exemplo, um *chatbot* descrito por (32) engaja-se em diálogos empáticos com os usuários, aliviando sentimentos de ansiedade e depressão ao responder com interações reconfortantes. De forma semelhante, o Coral (100) foi projetado para atuar como um agente de conversa empático, envolvendo os usuários em diálogos de múltiplos turnos em domínio aberto para criar uma atmosfera de suporte. Esses *chatbots* se concentram no bem-estar emocional, proporcionando aos usuários a sensação de serem compreendidos e cuidados em momentos de angústia.

Além do apoio emocional, muitos *chatbots* são projetados para *Treinamento Mo-*

tivacional e de Comportamento. Esses *chatbots* ajudam os usuários a definir e alcançar metas pessoais, frequentemente utilizando técnicas de mudança comportamental para incentivar estilos de vida ou hábitos mais saudáveis. Por exemplo, o CoachAI (29) é projetado para auxiliar no coaching de saúde, promovendo atividade física, dietas saudáveis e gerenciamento de estresse por meio de intervenções personalizadas. O *chatbot* conduz conversas estruturadas que motivam os usuários a adotar comportamentos positivos e fazer mudanças de estilo de vida sustentáveis. No domínio da cessação do tabagismo, um *chatbot* descrito por (5) emula a EM para ajudar fumantes a refletirem sobre seu comportamento e se moverem em direção a parar de fumar, guiando-os por um processo motivacional estruturado e baseado em dados.

Em *Ferramentas de Treinamento de Terapeutas*, *chatbots* ajudam a treinar terapeutas e profissionais de saúde mental. Por exemplo, o ClientBot (112) foi projetado para simular um paciente em interações baseadas em texto, fornecendo *feedback* em tempo real para terapeutas em treinamento. Este *chatbot* ajuda os terapeutas a desenvolver suas habilidades de aconselhamento ao simular uma conversa terapêutica realista, como fazer perguntas abertas e fornecer reflexões. Essas ferramentas desempenham um papel essencial no treinamento profissional, permitindo que os terapeutas pratiquem e aperfeiçoem suas habilidades em um ambiente virtual controlado.

Por fim, os *chatbots* também estão sendo usados para *Facilitação de Suporte entre Pares*, onde ajudam os usuários a se engajar em redes de suporte entre pares. O KokoBot (79) é um excelente exemplo de um *chatbot* projetado para facilitar interações entre pares ao ensinar habilidades de reavaliação cognitiva e simular respostas empáticas. Ao promover um ambiente onde os usuários podem aprender uns com os outros e oferecer suporte emocional, o KokoBot incentiva um senso de comunidade e compreensão compartilhada entre seus usuários. Outro exemplo é um *chatbot* descrito por (31), que aumenta o engajamento em grupos de suporte online de saúde ao detectar intenções dos usuários e responder com conteúdo relevante e baseado em evidências. Isso ajuda a melhorar a qualidade das discussões nesses grupos, garantindo que os usuários recebam informações úteis e precisas.

2.2 PROTOCOLO DA REVISÃO SISTEMÁTICA

O processo metodológico foi meticulosamente planejado e implementado seguindo as diretrizes estabelecidas por (90), garantindo uma análise abrangente e rigorosa. A metodologia está dividida nas seguintes etapas: (i) formulação de questões de pesquisa, (ii) identificação de termos de busca pertinentes, (iii) estabelecimento de critérios de exclusão e (iv) escolha de bases de dados apropriadas para a obtenção de literatura. Cada etapa é crucial para alcançar os objetivos da pesquisa e compreender as tendências e inovações atuais no uso de *chatbots* para o bem-estar mental.

2.2.1 Foco da Revisão

Esta revisão sistemática foca em *chatbots* textuais na área de saúde mental, destacando sua acessibilidade. *Chatbots* textuais oferecem um acesso mais amplo, incluindo usuários de dispositivos básicos e regiões com infraestrutura tecnológica limitada (2). Além disso, possibilitam o uso anônimo, o que pode mitigar barreiras para buscar ajuda psicológica (51). Essa escolha está alinhada com o potencial dos *chatbots* textuais de oferecer suporte contínuo e discreto, o que é crucial para indivíduos enfrentando desafios de saúde mental (30).

2.2.2 Período e Significado

Foram analisados 84 artigos publicados entre janeiro de 2017 e 19 de fevereiro de 2024, período escolhido devido ao notável desenvolvimento e interesse em chatbots. Esse intervalo corresponde a avanços significativos em inteligência artificial e processamento de linguagem natural aplicados à saúde mental, destacando a crescente importância dos *chatbots* textuais como uma ferramenta de apoio (22, 65).

2.2.3 String de busca

A formulação da string de busca foi orientada pelo método PICOC, uma estratégia reconhecida para a formulação de consultas em revisões sistemáticas (72). A configuração de nossa string de busca é mostrada na Tabela 1.

Tabela 1 – Configuração da String de Busca.

Elemento PICOC	**Palavras-chave principais**	**Expressão de busca**
População	*Chatbots*	*(“conversational agent*” OR “relational agent*” OR “chatbot*” “OR “virtual assistant*” OR “virtual agent” OR ”natural language dialogue” OR “dialogue system”)*
Intervenção	-	-
Comparação	-	-
Resultado	Abordagens e tecnologias	*(“approache*” OR ”method*” OR “techn*” OR “strategie*” OR “algorithm*” OR “tool” OR “framework*” OR “model*” OR “develop*” OR “building*” OR “system*”)*
Contexto	Saúde mental	*(“mental health*” OR “mental illness*” OR “mental disorder*” OR “mental state*” OR “psycholog*” OR “health coaching*” OR “psychiat*” OR “therap*”)*

Fonte: Elaborado pelo autor (2024).

Utilizamos Scopus, PubMed e Google Scholar para esta revisão sistemática. Esses bancos de dados foram selecionados com base em suas forças em fornecer uma base sólida e diversificada, crucial para entender as tendências atuais e as inovações no uso de *chatbots* para o bem-estar mental. A combinação de Scopus, PubMed e Google Scholar assegura uma abordagem abrangente e interdisciplinar, capturando pesquisas revisadas por pares de alta qualidade e literatura cinzenta que podem oferecer *insights* únicos sobre o tema.

2.2.4 Coleta, Inclusão e Exclusão de Artigos

Inicialmente, realizamos uma busca automatizada no Google Scholar utilizando um algoritmo sistemático para executar a string de busca para cada ano, de 2017 a fevereiro de 2024. Essa abordagem foi necessária devido à limitação do Google Scholar de exibir apenas os primeiros 1.000 resultados para cada consulta. Adicionalmente, as buscas foram realizadas no PubMed e no Scopus, inserindo diretamente a string de busca. Para validar a eficácia da string de busca, verificamos com uma lista de referência contendo seis artigos-chave. Após confirmar que todos os artigos de referência foram capturados, prosseguimos com a coleta de artigos.

A identificação inicial resultou em 4.542 artigos, reduzidos para 2.947 após a remoção de duplicatas. Foram recuperados 1.563 artigos do Google Scholar, 1.165 do Scopus e 219 do PubMed.

Estabelecemos critérios claros de inclusão e exclusão para refinar o processo de seleção de artigos:

Critérios de Inclusão:

- O estudo foca em *chatbots* textuais na saúde mental.
- O artigo está escrito em inglês, português ou espanhol.
- O texto completo do manuscrito está disponível.

Critérios de Exclusão:

- O estudo não foca em *chatbots* textuais na saúde mental.
- O artigo não está escrito em inglês, português ou espanhol.
- O texto completo do manuscrito não está disponível.
- O modelo ou as tecnologias utilizadas no *chatbot* não são descritos.

Realizamos um processo sistemático de triagem. Inicialmente, os títulos foram revisados, resultando em 460 artigos. Subsequentemente, os resumos foram avaliados, restando 131 artigos. Finalmente, os artigos foram lidos integralmente, levando a 104

artigos. Quarenta artigos foram excluídos devido à falta de informações suficientes sobre os modelos utilizados, totalizando 84 artigos elegíveis para análise.

2.2.5 Visão Geral do Conjunto de Dados Final

Nesta subseção, apresentamos uma visão abrangente dos dados dos 84 estudos coletados neste trabalho. O número de estudos publicados por ano aumentou de forma constante, refletindo o crescente interesse em *chatbots* para a saúde mental. Um aumento notável foi observado, particularmente a partir de 2020, com um pico em 2023, quando 25 estudos foram publicados (ver Tabela 2). Essa tendência destaca a crescente atenção para esse tópico dentro da comunidade de pesquisa. Nossa revisão coletou dados até o início de 2024, e a tendência sugere que o número de estudos que avaliam ou implementam essas ferramentas continuará a crescer.

Tabela 2 – Número de estudos publicados por ano.

Ano	**Quantidade**	**Percentual (%)**
2024*	11	13.10
2023	25	29.76
2022	20	23.81
2021	13	15.48
2020	8	9.52
2019	4	4.76
2018	1	1.19
2017	2	2.38

Fonte: Elaborado pelo autor (2024).

Em seguida, analisamos os papéis desempenhados pelos chatbots. Através de uma análise do conjunto final, identificamos quatro grandes categorias nas quais as funções desses sistemas podem ser classificadas: I. Abordagens Psicoterapêuticas, II. Intervenções de Suporte e Educacionais, III. Ferramentas de Avaliação e Diagnóstico e IV. Técnicas de Comunicação e Aconselhamento (ver Tabela 3). Cada *chatbot* dentro do conjunto de dados foi projetado para cumprir pelo menos uma dessas funções, com alguns desempenhando múltiplas funções, refletindo a natureza multifacetada das intervenções em saúde mental via agentes conversacionais.

As Abordagens Psicoterapêuticas emergem como o papel dominante, compreendendo quase metade de todos os *chatbots* revisados. Esses sistemas são projetados para replicar ou complementar sessões terapêuticas tradicionais. Por exemplo, (119) utiliza a Terapia de Solução de Problemas (PST) para auxiliar cuidadores, enquanto (34) empregam técnicas de estimulação cognitiva para ajudar usuários idosos, ilustrando a adaptabilidade desses *chatbots* para diferentes populações. Notavelmente, RehabChat (46) visa apoiar

Tabela 3 – Papéis desempenhados pelos chatbots.

Categoria	**Quantidade**	**Percentual (%)**
I. Abordagens Psicoterapêuticas	41	48.81
II. Intervenções de Suporte e Educacionais	33	39.29
III. Ferramentas de Avaliação e Diagnóstico	17	20.24
IV. Técnicas de Comunicação e Aconselhamento	7	8.33

Fonte: Elaborado pelo autor (2024).

indivíduos em recuperação de lesões cerebrais, oferecendo suporte para definição de metas e motivação, demonstrando o amplo potencial de aplicações dos *chatbots* psicoterapêuticos.

Estudos adicionais que exploram *chatbots* no contexto de Abordagens Psicoterapêuticas incluem trabalhos de (84, 68, 59, 62, 12, 111, 116, 34, 14, 95, 110, 45, 17, 9, 39, 38).

As Intervenções de Suporte e Educacionais, o segundo papel mais comum, fornecem aos usuários recomendações de saúde mental e recursos psicoeducacionais. Sistemas como o TherapyBot (106) ajudam a gerenciar a ansiedade por meio de conversas estruturadas, enquanto outros, como (87), oferecem conselhos de saúde mental via estruturas de correspondência de intenções pré-definidas. Essas intervenções geralmente são projetadas para guiar os usuários em estratégias de gerenciamento de estresse e enfrentamento, contribuindo para seu bem-estar geral.

Exemplos adicionais de *chatbots* envolvidos em Intervenções de Suporte e Educacionais incluem trabalhos de (35, 85, 37, 86, 94, 78, 1, 128, 24, 91).

A terceira categoria, Ferramentas de Avaliação e Diagnóstico, foca em rastrear e diagnosticar condições de saúde mental, frequentemente utilizando escalas clinicamente validadas. Por exemplo, (124) introduzem um *chatbot* que realiza triagens de depressão e avaliações emocionais, enquanto (57) empregam um processo de entrevista estruturado para detectar sinais precoces de depressão. Esses sistemas são fundamentais para expandir o diagnóstico em saúde mental para formatos digitais acessíveis e escaláveis, particularmente úteis para intervenções precoces.

Outros exemplos incluem (61, 101, 71, 80, 74, 67).

Por fim, as Técnicas de Comunicação e Aconselhamento englobam *chatbots* que facilitam a expressão do usuário e o compartilhamento emocional. Sistemas como os descritos por (64) visam melhorar a comunicação dos usuários com profissionais de saúde mental ou atuam como mediadores, criando um espaço seguro para que os usuários expressem seus pensamentos e sentimentos.

Outros exemplos incluem (64, 58, 76).

Curiosamente, alguns *chatbots* desempenham múltiplos papéis. OkBot (52), por exemplo, combina tanto Abordagens Psicoterapêuticas quanto Ferramentas de Avaliação,

oferecendo processos de terapia guiada e avaliações de rastreamento de humor em bahasa malaia, expandindo o acesso aos recursos de saúde mental para populações não falantes de inglês.

Paralelamente aos papéis funcionais desses sistemas, um conjunto diversificado de fundamentos teóricos sustenta o *design* dos chatbots. A Tabela 4 resume a distribuição desses fundamentos nos estudos. A Terapia Cognitivo-Comportamental (TCC) e suas extensões são os fundamentos mais comumente empregados, presentes em 25% dos *chatbots* revisados. Isso se alinha com a proeminência da TCC em intervenções psicoterapêuticas baseadas em evidências para condições de saúde mental. Técnicas de Suporte Emocional, Estratégias Psicoeducacionais e Ferramentas de Análise de Sentimento e Emoção também figuram de forma significativa, destacando uma ampla dependência de metodologias psicológicas estabelecidas para garantir a eficácia nas intervenções em saúde mental (30, 89).

Tabela 4 – Fundamentos Teóricos dos Chatbots.

Fundamento Teórico	**Quantidade**	**Percentual (%)**
TCC e Extensões	21	25.00
Técnicas de Suporte Emocional	15	17.86
Estratégias Psicoeducacionais	13	15.48
Ferramentas de Análise de Emoção/Sentimento	10	11.90
MI e Suporte	10	11.90
Avaliações Padronizadas	7	8.33
Ativação e Reforço Comportamental	5	5.95
Estratégias de Aconselhamento	5	5.95
Sugestões de Mudança de Estilo de Vida	5	5.95
Intervenções de Psicologia Positiva	4	4.76
IPT	3	3.57
Gerenciamento de Estresse e Ansiedade	3	3.57
Técnicas de Intervenção em Crise	2	2.38
Terapia Dialética Comportamental (DBT)	2	2.38
Práticas de Mindfulness	2	2.38
Práticas Reflexivas	2	2.38
Técnicas de Solução de Problemas e Cognitivas	2	2.38
Terapia de Aceitação e Compromisso (ACT)	1	1.19
Estratégias de Suporte para Vícios	1	1.19
Terapia Centrada na Pessoa (TCP)	1	1.19
Avaliações de Personalidade	1	1.19

Fonte: Elaborado pelo autor (2024).

Quanto às condições de saúde mental focadas por esses chatbots, observamos que depressão e ansiedade dominam o cenário (ver Tabela 5), refletindo a prevalência dessas condições dentro da comunidade de saúde mental. No entanto, uma proporção considerável dos *chatbots* foi projetada para abordar questões gerais de saúde mental, indicando uma

tendência para o desenvolvimento de sistemas mais holísticos que atendem a um espectro amplo de necessidades de saúde mental.

Tabela 5 – Focos de Saúde Mental dos Chatbots.

Foco de Saúde Mental	**Quantidade**	**Percentual (%)**
Problemas de saúde mental	36	42.86
Depressão	21	25.00
Ansiedade	13	15.48
Angústia geral	13	15.48
Humor	3	3.57

Fonte: Elaborado pelo autor (2024).

Equipados com um papel específico, fundamento teórico e foco de saúde mental, diversos componentes de modelos de linguagem sustentam esses chatbots. Em relação aos tipos de componentes de modelos aplicados a *chatbots* de saúde mental, observamos a distribuição descrita na Tabela 6. Cada componente possui particularidades, capacidades associadas e limitações inerentes. Por exemplo, *chatbots* com um componente gerativo apresentam maior flexibilidade de comunicação, permitindo respostas mais naturais e variadas. No entanto, podem também dificultar a modelagem de intervenções estruturadas dentro de um quadro bem definido, o que pode representar desafios para garantir a segurança e a adequação da interação. Em contraste, *chatbots* que dependem de uma lógica baseada em regras podem oferecer interações mais controladas e previsíveis, mas podem carecer da adaptabilidade dos modelos gerativos (70).

Tabela 6 – Componentes de Modelos Usados em Chatbots.

Componente de Modelo	**Quantidade**	**Percentual (%)**
Lógica baseada em regras	48	57.14
Classificação de intenções	44	52.38
Componente gerativo	31	36.90
Reconhecimento de emoção/sentimento	18	21.43
Recuperação de conhecimento	16	19.05

Fonte: Elaborado pelo autor (2024).

Neste contexto, começamos a situar o objeto de interesse das nossas perguntas de pesquisa: Existe um grau de associação entre os componentes utilizados pelos modelos e seus respectivos papéis? Modelos que visam problemas específicos de saúde mental estão mais ou menos associados a diferentes componentes? Essas considerações são cruciais para entender como o *design* dos componentes do *chatbot* influencia sua eficácia e adequação para diversas intervenções em saúde mental (7).

Por fim, analisamos os atributos avaliados em cada sistema de *chatbot* (ver Tabela 7). Os atributos mais frequentemente avaliados foram o desempenho do sistema (tanto

intrínseco quanto extrínseco), refletindo a preocupação central de garantir que esses *chatbots* funcionem de forma confiável e apresentem resultados consistentes. Atributos como resultados em saúde mental, satisfação do usuário e confiança também foram comumente avaliados, enfatizando a necessidade de eficácia técnica e experiências positivas para o usuário.

Tabela 7 – Atributos Avaliados nos Chatbots.

Atributo	**Quantidade**	**Percentual (%)**
Resultados em saúde mental/comportamentais	29	34.52
Confiança e construção de relacionamento	7	8.33
Engajamento e adesão do usuário	7	8.33
Satisfação e *feedback* do usuário	22	26.19
Desempenho do sistema (Extrínseco)	24	28.57
Desempenho do sistema (Intrínseco)	35	41.67

Fonte: Elaborado pelo autor (2024).

Atributos focados no usuário, como resultados em saúde mental e satisfação do usuário, são críticos para avaliar o impacto desses *chatbots* no mundo real. Eles refletem o quão bem esses sistemas são percebidos pelos usuários e até que ponto promovem o engajamento e a adesão às intervenções terapêuticas. Por outro lado, as métricas de desempenho do sistema fornecem *insights* sobre as capacidades técnicas dos chatbots, incluindo confiabilidade, precisão e eficiência.

Em suma, os dados desses 84 estudos apresentam uma visão abrangente do panorama em evolução das aplicações de *chatbots* em saúde mental. As tendências sugerem uma sofisticação crescente tanto no *design* quanto na avaliação desses sistemas, com um foco crescente na integração de fundamentos terapêuticos robustos, no atendimento de necessidades críticas de saúde mental e na entrega de resultados centrados no usuário por meio de soluções tecnológicas bem fundamentadas.

2.3 RQ1: COMO OS MODELOS DE *CHATBOT* VARIAM NA IMPLEMENTAÇÃO PARA DIFERENTES CASOS DE USO?

Nesta seção, exploraremos a relação entre o *design* das arquiteturas de modelos conversacionais e seus respectivos casos de uso, visando extrair *insights*, tendências potenciais e oportunidades de pesquisa. Primeiro, examinaremos a relação entre a variável **Focos de Saúde Mental** (relacionados aos cenários de uso) e **Componentes do Modelo de Chatbot** (que abrangem as características intrínsecas das tecnologias utilizadas nos modelos). Em seguida, investigaremos a relação entre **Papéis dos Chatbots** e **Componentes do Modelo de Chatbot** para entender como diferentes papéis influenciam o *design* e a funcionalidade dos modelos de chatbot.

2.3.1 Componentes do Modelo de *Chatbot* e Focos de Saúde Mental

Para caracterizar as relações e padrões de interesse, analisamos os resultados da tabela de contingência (Tabela 8) e da tabela de resíduos padronizados do qui-quadrado (Tabela 9), que fornecem uma compreensão detalhada das relações entre os componentes do modelo de *chatbot* e os focos de saúde mental. Utilizando a correção de Bonferroni para ajustar para múltiplos testes de significância, o valor crítico ajustado para considerar uma associação significativa é aproximadamente 2.63, assumindo um nível de significância de 0.05 e um total de 30 comparações ($\alpha' = \frac{0.05}{30}$).

Os resíduos padronizados indicam algumas associações notáveis. O componente de Recuperação de Conhecimento mostra um valor positivo significativo em relação ao foco Bem-estar Psicológico (2.3643), que, embora não alcance o limiar crítico ajustado de 2.63, está muito próximo, sugerindo uma associação positiva relevante. Esse resultado implica que a recuperação de conhecimento está mais associada à promoção do bem-estar psicológico do que o esperado sob a hipótese de independência. Em contraste, os resíduos negativos para Ansiedade (-1.4275), Angústia Geral (-1.3736) e Humor (-0.6729) indicam que a Recuperação de Conhecimento é menos frequentemente utilizada ou é menos eficaz em contextos focados nesses alvos específicos de saúde mental.

O componente Lógica Baseada em Regras mostra um resíduo positivo para Angústia Geral (1.1025), que, apesar de não ser significativo após a correção de Bonferroni, destaca uma tendência de associação positiva. Isso sugere que abordagens baseadas em regras têm uma modesta relação com o gerenciamento da angústia geral. Enquanto isso, o componente Classificação de Intenções apresenta um resíduo positivo moderado para Humor (0.9401), indicando uma associação mais forte do que o esperado, embora ainda dentro do intervalo não significativo. Esses resultados sugerem que a classificação de intenções pode estar relativamente mais envolvida em interações relacionadas ao humor.

Outros componentes não apresentaram resíduos significativos. Reconhecimento de Emoção/Sentimento não mostra fortes associações com focos de saúde mental, com valores de resíduos próximos de zero, sugerindo uma dispersão uniforme sem uma tendência clara de sobre ou sub-representação. Isso indica que sua aplicação não se destaca em relação a nenhum foco específico, apesar da utilidade intuitiva do reconhecimento de emoções e sentimentos. Em contrapartida, o Componente Gerativo mostra um valor relativamente alto para Problemas de Saúde Mental (0.8798), mas sem alcançar significância estatística. Essa dispersão de resíduos reflete uma aplicação ampla e possivelmente menos direcionada desse tipo de tecnologia em vários contextos de saúde mental.

De uma perspectiva crítica, a análise revela que a maioria das associações entre componentes do modelo e focos de saúde mental não atinge significância após a correção de Bonferroni, destacando a complexidade inerente dessas interações. Isso sugere que modelos que dependem exclusivamente de uma única abordagem, como lógica baseada em regras

ou recuperação de conhecimento, podem não ser suficientes para capturar a diversidade de necessidades em diferentes contextos de saúde mental. A flexibilidade de alguns componentes, como Classificação de Intenções e Reconhecimento de Emoção/Sentimento, permite seu uso em uma ampla gama de aplicações, mas essa mesma generalização pode limitar sua eficácia em proporcionar intervenções mais direcionadas.

Esses resultados enfatizam a importância de combinar diferentes componentes para criar abordagens híbridas que atendam melhor às necessidades específicas dos focos de saúde mental. Além disso, destacam a necessidade de uma maior personalização das intervenções, ajustando modelos para maximizar seu impacto nos contextos desejados. A análise crítica sugere que, embora os componentes individualmente ofereçam valor, seu potencial completo será alcançado apenas pela integração de técnicas que abordem as nuances dos estados de saúde mental, promovendo intervenções mais eficazes e contextualizadas.

Tabela 8 – Tabela de Contingência para Componentes de Modelo e Focos de Saúde Mental.

Componente do Modelo	**Ansiedade**	**Depressão**	**Angústia Geral**	**Problemas de Saúde Mental**	**Humor**	**Bem-estar Psicológico**
Reconhecimento de Emoção/Sentimento	4	7	4	7	1	5
Componente Gerativo	4	8	3	15	1	8
Classificação de Intenções	9	14	7	20	3	9
Recuperação de Conhecimento	0	2	0	6	0	8
Lógica Baseada em Regras	10	11	11	17	1	17

Fonte: Elaborado pelo autor (2024).

A Análise de Correspondência (AC) apresentada na Figura 1 fornece uma visão gráfica das relações entre componentes de modelo de *chatbot* e focos de saúde mental, complementando a análise dos resíduos padronizados. A AC transforma as associações em um espaço bidimensional, facilitando a interpretação das conexões entre categorias e permitindo a visualização de padrões latentes. A Tabela 10 mostra que as duas primeiras dimensões juntas explicam 96.19% da variabilidade dos dados, com a *Dimension 1* contribuindo com 68.82% e a *Dimension 2* com 27.37%. Essa alta variabilidade explicada indica que o mapa perceptual gerado pela AC captura de forma robusta as associações significativas entre os componentes e focos analisados. A inércia associada a cada dimensão

Tabela 9 – Tabela de Resíduos Padronizados para Componentes de Modelo e Focos de Saúde Mental.

Componente do Modelo	**Ansiedade**	**Depressão**	**Angústia Geral**	**Problemas de Saúde Mental**	**Humor**	**Bem-estar Psicológico**
Reconhecimento de Emoção/Sentimento	0.2298	0.6168	0.3842	-0.5409	0.2331	-0.4847
Componente Gerativo	-0.4339	0.0984	-0.7456	0.8798	-0.0988	-0.2198
Classificação de Intenções	0.3928	0.4899	-0.1151	0.2272	0.9401	-1.2799
Recuperação de Conhecimento	-1.4275	-0.6571	-1.3736	0.4941	-0.6729	2.3643
Lógica Baseada em Regras	0.5022	-0.6240	1.1025	-0.7816	-0.6508	0.5569

Fonte: Elaborado pelo autor (2024).

é crucial para entender a força das associações capturadas pela análise. A alta contribuição da *Dimension 1* para a inércia total sugere que a maior parte da variabilidade observada está associada a um eixo que separa fortemente componentes e focos ao longo dessa dimensão. Por outro lado, a *Dimension 2* adiciona uma camada adicional de variabilidade, capturando associações que, embora menos dominantes, ainda são significativas. Essa estrutura indica que a AC é altamente eficiente em capturar relações críticas, com uma inércia residual mínima (3.81%), reforçando que poucas associações relevantes permanecem não explicadas.

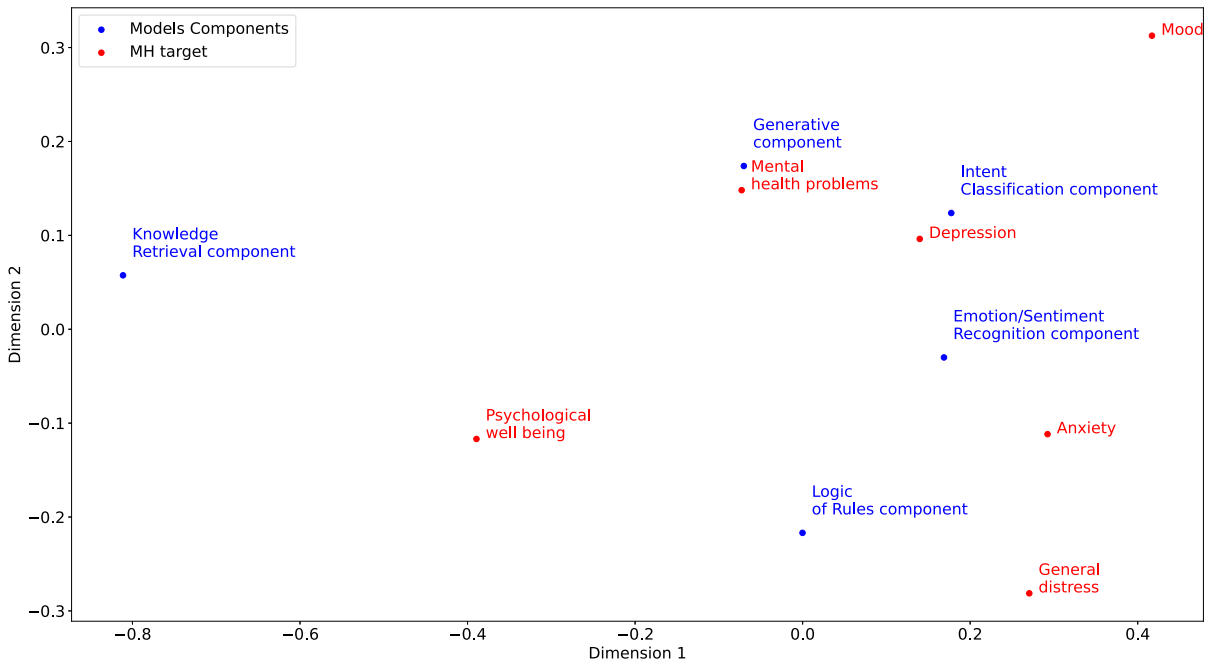
Tabela 10 – Variância Explicada pelas Dimensões

Dimensão	**Variância Explicada (%)**
Dimension 1	68.82
Dimension 2	27.37

Fonte: Elaborado pelo autor (2024).

As coordenadas das linhas dos componentes do modelo, apresentadas na Tabela 11, mostram como cada componente está posicionado em relação às duas dimensões. A Recuperação de Conhecimento se destaca como o componente mais distante na *Dimension 1*, com uma coordenada de -0.8112, sugerindo uma forte associação com focos posicionados negativamente ao longo dessa dimensão, especificamente Bem-estar Psicológico (-0.3894), conforme mostrado na Tabela 12. Essa relação foi sugerida anteriormente pela análise

Figura 1 – Análise de Correspondência (AC) para Componentes de Modelo vs Focos de Saúde Mental.



Fonte: Elaborado pelo autor (2024).

dos resíduos padronizados, onde a Recuperação de Conhecimento mostrou um resíduo positivo para Bem-estar Psicológico, indicando uma conexão relevante agora visualmente confirmada pela AC.

Tabela 11 – Coordenadas das Linhas (Componentes do Modelo)

Componente do Modelo	**Dim 1**	**Dim 2**
Reconhecimento de Emoção/Sentimento	0.1687	-0.0301
Componente Gerativo	-0.0702	0.1738
Classificação de Intenções	0.1774	0.1238
Recuperação de Conhecimento	-0.8112	0.0574
Lógica Baseada em Regras	-0.0001	-0.2169

Fonte: Elaborado pelo autor (2024).

Por outro lado, a Lógica Baseada em Regras está fortemente associada à *Dimension 2* (-0.2169), contribuindo substancialmente para essa dimensão, conforme evidenciado na Tabela 13 com uma contribuição de 0.5882, a maior entre todos os componentes nesta dimensão. Isso sugere que a Lógica Baseada em Regras possui um perfil distinto, estando mais fortemente associada a focos que também contribuem significativamente para a *Dimension 2*, como Angústia Geral (0.3692) e Problemas de Saúde Mental (0.2661). Esse comportamento é consistente com os resíduos padronizados, que sugeriram uma associação positiva entre a Lógica Baseada em Regras e Angústia Geral, reforçando a interpretação de

Tabela 12 – Coordenadas das Colunas (Focos de Saúde Mental)

Foco de Saúde Mental	**Dim 1**	**Dim 2**
Ansiedade	0.2924	-0.1118
Depressão	0.1398	0.0962
Angústia Geral	0.2706	-0.2813
Problemas de Saúde Mental	-0.0728	0.1481
Humor	0.4170	0.3127
Bem-estar Psicológico	-0.3894	-0.1169

Fonte: Elaborado pelo autor (2024).

que abordagens baseadas em regras podem ter uma influência moderada nesse contexto.

Tabela 13 – Contribuições dos Componentes do Modelo para as Dimensões

Componente do Modelo	**Contribuição Dim 1**	**Contribuição Dim 2**
Reconhecimento de Emoção/Sentimento	0.0592	0.0047
Componente Gerativo	0.0143	0.2200
Classificação de Intenções	0.1449	0.1773
Recuperação de Conhecimento	0.7816	0.0098
Lógica Baseada em Regras	0.0000	0.5882

Fonte: Elaborado pelo autor (2024).

A Classificação de Intenções, com coordenadas de 0.1774 e 0.1238 nas Dimensões 1 e 2, respectivamente, aparece centralizada no espaço bidimensional, sugerindo uma associação moderada e dispersa entre vários focos, sem uma conexão dominante com nenhum foco específico. Suas contribuições para as dimensões são notáveis (0.1449 e 0.1773), destacando sua flexibilidade e presença em diferentes contextos. Isso é consistente com a análise de resíduos, que indicou uma leve associação com Humor. Esse foco também apresenta coordenadas positivas em ambas as dimensões, mas não se destaca fortemente em nenhuma delas.

O Reconhecimento de Emoção/Sentimento apresenta coordenadas relativamente neutras em ambas as dimensões (0.1687 e -0.0301), refletindo sua posição próxima ao centro da análise, sugerindo um uso amplo, mas não específico. As contribuições modestas para as dimensões (0.0592 e 0.0047) indicam que, embora amplamente aplicável, esse componente não é particularmente influente em nenhuma dimensão específica, corroborando os resultados dos resíduos padronizados que não indicaram associações fortes com nenhum foco de saúde mental específico.

O Componente Gerativo, com coordenadas de -0.0702 na *Dimension 1* e 0.1738 na *Dimension 2*, mostra uma leve inclinação para os focos associados positivamente à *Dimension 2*, como Problemas de Saúde Mental e Humor. Sua contribuição relativamente

alta para a *Dimension 2* (0.2200) sugere que este componente pode ser mais relevante em contextos focados em humor e problemas gerais de saúde mental. No entanto, essa relação não foi estatisticamente forte nos resíduos padronizados.

A análise das contribuições dos focos de saúde mental para as dimensões, conforme apresentado na Tabela 14, mostra que Bem-estar Psicológico é o foco que mais contribui para a *Dimension 1* (0.5289), destacando sua forte associação com componentes posicionados negativamente nessa dimensão, particularmente a Recuperação de Conhecimento. Essa alta contribuição reforça a percepção de que estratégias baseadas em recuperação de conhecimento podem estar significativamente ligadas à promoção do bem-estar psicológico, embora essa associação não tenha atingido significância crítica nos resíduos padronizados após a correção de Bonferroni.

Tabela 14 – Contribuições dos Focos de Saúde Mental para as Dimensões

Foco de Saúde Mental	**Contribuição Dim 1**	**Contribuição Dim 2**
Ansiedade	0.1713	0.0629
Depressão	0.0609	0.0725
Angústia Geral	0.1359	0.3692
Problemas de Saúde Mental	0.0256	0.2661
Humor	0.0775	0.1095
Bem-estar Psicológico	0.5289	0.1198

Fonte: Elaborado pelo autor (2024).

Por outro lado, Angústia Geral é o foco que mais contribui para a *Dimension 2* (0.3692), sugerindo que essa dimensão captura um eixo de variação relacionado ao gerenciamento de estados de angústia, onde Lógica Baseada em Regras e, em menor grau, Classificação de Intenções, desempenham papéis notáveis. A associação da *Dimension 2* com Angústia Geral também aponta para a necessidade de explorar como as abordagens baseadas em regras e classificação de intenções podem ser ajustadas para abordar melhor esses estados emocionais.

Criticamente, a Análise de Correspondência sublinha a necessidade de uma abordagem multifacetada, já que a maioria dos componentes apresenta uma dispersão de associações sem um alinhamento extremamente forte com um único foco de saúde mental. Isso reflete a complexidade dos estados mentais humanos e a necessidade de combinações integradas de técnicas para maximizar a eficácia da intervenção. Além disso, a grande variabilidade explicada pelas duas primeiras dimensões sugere que, embora existam padrões discerníveis, muitos componentes e focos exibem interações complexas que uma única dimensão de análise não pode capturar completamente.

Esses resultados destacam que, embora existam associações relevantes entre certos componentes e focos, como o vínculo entre Recuperação de Conhecimento e Bem-estar

Psicológico ou Lógica Baseada em Regras e Angústia Geral, a aplicação eficaz desses componentes em intervenções de saúde mental requer um ajuste cuidadoso e a consideração de múltiplas técnicas para atender à diversidade de necessidades individuais. A análise aponta para potenciais melhorias na personalização e na combinação estratégica de componentes para abordar as nuances de diferentes focos de saúde mental.

2.3.2 Componente do Modelo de *Chatbot* e Papéis

Continuando a entender as possíveis relações entre os modelos usados em *chatbots* e suas aplicações práticas, conduzimos agora uma análise comparando Componentes do Modelo e Papéis dos Chatbots. A análise dos componentes do modelo de *chatbot* em relação aos papéis que esses sistemas desempenham na saúde mental, conforme mostrado na Tabela 15, busca identificar padrões de associação entre as características técnicas dos *chatbots* e suas aplicações práticas. Usando a Tabela de Resíduos Padronizados (Tabela 16), examinamos a significância e a direção dessas associações, aplicando uma correção de Bonferroni para ajustar a significância estatística para múltiplas comparações.

Tabela 15 – Tabela de Contingência dos Componentes do Modelo e Papéis

Componente do Modelo	**I. Abordagens Psicoterapêuticas**	**II. Intervenções de Suporte e Educacionais**	**III. Ferramentas de Avaliação e Diagnóstico**	**IV. Técnicas de Comunicação e Aconselhamento**
Reconhecimento de Emoção/Sentimento	10	3	9	0
Componente Gerativo	11	16	6	6
Classificação de Intenções	19	13	14	4
Recuperação de Conhecimento	6	9	2	1
Lógica Baseada em Regras	29	18	6	2

Fonte: Elaborado pelo autor (2024).

Os resíduos padronizados indicam desvios das frequências esperadas, e valores absolutos maiores que aproximadamente 2 sugerem associações significativas após a correção de Bonferroni, considerando o contexto de múltiplos testes. Primeiramente, observamos que o componente Reconhecimento de Emoção/Sentimento possui um resíduo significativamente elevado (2.1757) em associação com III. Ferramentas de Avaliação e Diagnóstico, indicando uma super-representação significativa dessa combinação. Essa

Tabela 16 – Tabela de Resíduos Padronizados para Componentes do Modelo e Papéis

Componente do Modelo	**I. Abordagens Psicoterapêuticas**	**II. Intervenções de Suporte e Educacionais**	**III. Ferramentas de Avaliação e Diagnóstico**	**IV. Técnicas de Comunicação e Aconselhamento**
Reconhecimento de Emoção/Sentimento	0.3448	-1.5265	2.1757	-1.2467
Componente Gerativo	-1.2282	0.9882	-0.6579	1.9546
Classificação de Intenções	-0.3058	-0.7574	1.2443	0.2487
Recuperação de Conhecimento	-0.4936	1.3437	-0.8513	-0.2410
Lógica Baseada em Regras	1.3900	0.0867	-1.5215	-0.9567

Fonte: Elaborado pelo autor (2024).

associação reflete a relevância do reconhecimento de emoções em avaliação e diagnóstico, onde a capacidade do *chatbot* de interpretar o estado emocional do usuário é essencial para fornecer um diagnóstico mais preciso e contextualizado. Por outro lado, o resíduo negativo com II. Intervenções de Suporte e Educacionais (-1.5265) sugere que essa combinação ocorre com menor frequência do que o esperado, embora não seja estatisticamente significativa após a correção, indicando uma tendência de baixa integração deste componente em intervenções de suporte e educação.

O Componente Gerativo mostra um resíduo positivo (1.9546) com IV. Técnicas de Comunicação e Aconselhamento, próximo ao limiar de significância, sugerindo que a geração de respostas abertas e criativas se alinha mais estreitamente com as técnicas de comunicação e aconselhamento, onde a flexibilidade de respostas é crucial. Em contraste, o resíduo negativo com I. Abordagens Psicoterapêuticas (-1.2282) indica uma leve sub-representação, possivelmente sugerindo que abordagens psicoterapêuticas preferem componentes mais estruturados, baseados em regras.

Para a Classificação de Intenções, os resíduos não indicam associações significativas após a correção de Bonferroni, com valores modestos em todas as categorias. No entanto, a associação com III. Ferramentas de Avaliação e Diagnóstico (1.2443) sugere que, embora não significativa, há uma leve tendência para o uso dessa técnica em contextos de avaliação, refletindo a importância de identificar corretamente as intenções dos usuários para uma triagem inicial de sintomas.

O componente Recuperação de Conhecimento mostra um resíduo positivo moderado com II. Intervenções de Suporte e Educacionais (1.3437), sugerindo uma tendência de uso de recuperação de conhecimento em contextos educacionais, onde a capacidade de fornecer informações baseadas em evidências é valiosa. Em contraste, a sub-representação com III. Ferramentas de Avaliação e Diagnóstico (-0.8513) e IV. Técnicas de Comunicação e Aconselhamento (-0.2410) reforça a ideia de que este componente é menos aplicável quando o foco está em avaliações e comunicação interpessoal.

Finalmente, a Lógica Baseada em Regras exibe um resíduo positivo com I. Abordagens Psicoterapêuticas (1.3900), indicando uma forte ligação entre abordagens estruturadas baseadas em regras e práticas psicoterapêuticas. Isso sugere que modelos que seguem uma lógica formal e estruturada são mais adequados para apoiar intervenções psicoterapêuticas, onde consistência e previsibilidade são essenciais. A associação negativa com III. Ferramentas de Avaliação e Diagnóstico (-1.5215) aponta para uma subutilização em contextos de avaliação, possivelmente devido à rigidez das regras que podem limitar a flexibilidade necessária para diagnósticos complexos.

Em resumo, a análise dos resíduos padronizados revela associações específicas e significativas entre componentes do modelo e papéis de chatbot. A correção de Bonferroni foi essencial para ajustar as significâncias devido ao número de comparações realizadas, permitindo uma interpretação crítica e estatisticamente robusta dos resultados. As associações detectadas refletem como diferentes componentes técnicos se alinham com as necessidades práticas dos *chatbots* em contextos de saúde mental, oferecendo *insights* para o *design* e aplicação otimizados desses sistemas em diversas intervenções.

A Análise de Correspondência (AC), ilustrada na Figura 2, visa explorar as relações entre os componentes dos modelos de *chatbot* e os papéis que esses sistemas desempenham na saúde mental. As tabelas de coordenadas de linhas e colunas (17 e 18) permitem identificar as posições relativas de cada componente do modelo e de cada papel de *chatbot* em um espaço dimensional reduzido, facilitando a visualização de associações e padrões de interação.

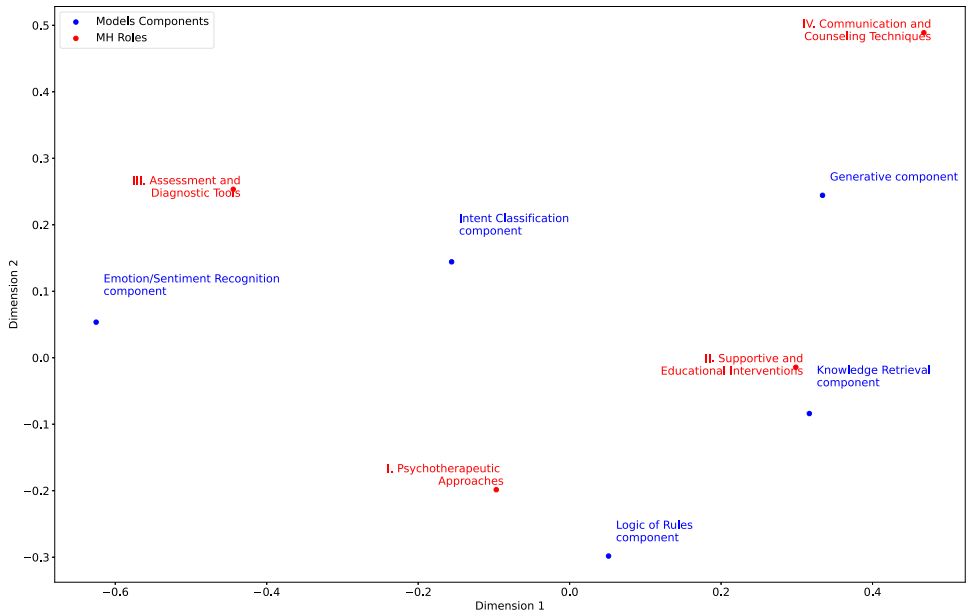
Tabela 17 – Coordenadas das Linhas dos Componentes do Modelo

Componente do Modelo	***Dimension 1***	***Dimension 2***
Reconhecimento de Emoção/Sentimento	−0.625289	0.053525
Componente Gerativo	0.333969	0.244289
Classificação de Intenções	−0.155858	0.144237
Recuperação de Conhecimento	0.316540	−0.083952
Lógica Baseada em Regras	0.051395	−0.298282

Fonte: Elaborado pelo autor (2024).

A Tabela 19 mostra que a *Dimension 1* explica 62.57% da variância, enquanto

Figura 2 – Análise de Correspondência (AC) para Componentes do Modelo vs Focos de Saúde Mental.



Fonte: Elaborado pelo autor (2024).

Tabela 18 – Coordenadas das Colunas dos Papéis dos Chatbots

Papel do Chatbot	***Dimension 1***	***Dimension 2***
I. Abordagens Psicoterapêuticas	−0.096903	−0.198526
II. Intervenções de Suporte e Educacionais	0.298699	−0.014265
III. Ferramentas de Avaliação e Diagnóstico	−0.444239	0.253379
IV. Técnicas de Comunicação e Aconselhamento	0.467793	0.488925

Fonte: Elaborado pelo autor (2024).

a *Dimension 2* captura 32.81%, totalizando 95.38% da variância explicada. Esse alto percentual indica que as duas dimensões são suficientes para capturar a maior parte da estrutura relacional entre os componentes e os papéis, permitindo uma interpretação robusta dos dados.

Tabela 19 – Variância Explicada pelas Dimensões

Dimensão	**Variância Explicada (%)**
Dimension 1	62.57
Dimension 2	32.81

Fonte: Elaborado pelo autor (2024).

As coordenadas dos componentes do modelo (Tabela 17) revelam que o Reconhecimento de Emoção/Sentimento tem uma forte associação negativa com a *Dimension 1* (-0.625289) e uma leve associação positiva com a *Dimension 2* (0.053525), sugerindo uma

relação distante de II. Intervenções de Suporte e Educacionais e proximidade com III. Ferramentas de Avaliação e Diagnóstico. Essa posição corrobora os resíduos padronizados observados anteriormente, onde o reconhecimento de emoção foi fortemente associado às ferramentas de avaliação, refletindo sua importância na identificação de estados emocionais críticos em contextos diagnósticos.

O Componente Gerativo, com coordenadas positivas em ambas as dimensões (0.333969 na *Dimension 1* e 0.244289 na *Dimension 2*), alinha-se fortemente com IV. Técnicas de Comunicação e Aconselhamento. Essa correspondência destaca a relevância dos componentes gerativos em contextos que exigem respostas dinâmicas e adaptativas, essenciais para técnicas de aconselhamento onde a flexibilidade de comunicação é fundamental.

A Classificação de Intenções aparece com coordenadas relativamente centrais (-0.155858 na *Dimension 1* e 0.144237 na *Dimension 2*), sugerindo uma relação equilibrada com os papéis, sem alinhamento claro com nenhum papel específico. Isso é refletido nos resíduos padronizados, que não mostraram associações significativas, indicando que a classificação de intenções é amplamente aplicável em vários contextos, mas sem uma predisposição marcada.

A Recuperação de Conhecimento está positivamente associada à *Dimension 1* (0.316540) e levemente negativa na *Dimension 2* (-0.083952), posicionando-se próxima a II. Intervenções de Suporte e Educacionais. Esse alinhamento reforça a ideia de que a recuperação de conhecimento é mais aplicada em contextos educacionais, onde o acesso a informações baseadas em evidências é vital para o suporte e a educação dos usuários.

A Lógica Baseada em Regras apresenta uma coordenada neutra na *Dimension 1* (0.051395) e uma forte associação negativa na *Dimension 2* (-0.298282), o que a alinha de perto com I. Abordagens Psicoterapêuticas. Esse alinhamento destaca a dependência das abordagens psicoterapêuticas em sistemas estruturados e regidos por regras, consistente com a aplicação de lógicas formais para garantir consistência e segurança nas intervenções terapêuticas.

Ao analisar as contribuições dos componentes para as dimensões (Tabela 20), o Reconhecimento de Emoção/Sentimento contribui significativamente para a *Dimension 1* (53.38%), reforçando seu papel central na variabilidade observada, particularmente em contextos de avaliação e diagnóstico. A Lógica Baseada em Regras contribui substancialmente para a *Dimension 2* (57.90%), destacando seu peso na estruturação de interações psicoterapêuticas, onde as regras desempenham um papel crítico.

Em relação aos papéis dos *chatbots* (Tabela 21), III. Ferramentas de Avaliação e Diagnóstico e IV. Técnicas de Comunicação e Aconselhamento têm altas contribuições para ambas as dimensões, sugerindo que essas aplicações desempenham papéis-chave na variabilidade dos dados. Isso reflete uma diversidade de abordagens, onde diferentes

Tabela 20 – Contribuições dos Componentes do Modelo para as Dimensões

Componente do Modelo	**Contribuição para Dim 1**	**Contribuição para Dim 2**
Reconhecimento de Emoção/Sentimento	53.38%	0.75%
Componente Gerativo	26.99%	27.54%
Classificação de Intenções	7.54%	12.31%
Recuperação de Conhecimento	11.19%	1.50%
Lógica Baseada em Regras	0.90%	57.90%

Fonte: Elaborado pelo autor (2024).

componentes do modelo influenciam as ferramentas de avaliação e diagnóstico e as técnicas de comunicação.

Tabela 21 – Contribuições dos Papéis dos *Chatbots* para as Dimensões

Papel do Chatbot	**Contribuição para Dim 1**	**Contribuição para Dim 2**
I. Abordagens Psicoterapêuticas	4.37%	34.98%
II. Intervenções de Suporte e Educacionais	32.67%	0.14%
III. Ferramentas de Avaliação e Diagnóstico	45.31%	28.11%
IV. Técnicas de Comunicação e Aconselhamento	17.65%	36.77%

Fonte: Elaborado pelo autor (2024).

A análise da inércia confirma que a maior parte da variabilidade nos dados é explicada pelas duas principais dimensões, com inércia total significativa, reforçando a validade das associações encontradas. As contribuições dos componentes e papéis indicam que há uma correspondência estrutural entre as características dos modelos e as demandas práticas dos chatbots, mostrando que os componentes técnicos são selecionados e desenvolvidos com base nas necessidades específicas de cada aplicação em saúde mental. Criticamente, embora a correspondência entre componentes e papéis forneça *insights* valiosos, é importante considerar as limitações das técnicas e a complexidade inerente do comportamento humano, que nem sempre pode ser capturada apenas por associações estatísticas. Portanto, o uso dessas ferramentas deve sempre ser complementado por uma avaliação contextual e uma compreensão profunda dos requisitos das intervenções psicológicas e de suporte.

2.3.3 Discussão

Dado o contexto das análises e testes anteriores, torna-se razoável considerar, embora nem sempre estatisticamente significativas para os nossos dados, que os modelos de *chatbot* variam em termos de tecnologias subjacentes e adequação para aplicações específicas em saúde mental. A visão geral de diversos sistemas de *chatbot* destaca como

diferentes componentes técnicos se alinham com os papéis práticos que esses sistemas desempenham, reforçando alguns dos *insights* derivados da Análise de Correspondência e das avaliações de resíduos padronizados.

Chatbots, como Woebot (30) e SELMA (44), projetados principalmente para Terapia Cognitivo-Comportamental (TCC), tipicamente dependem de sistemas baseados em regras e caminhos de conversação predefinidos. Esses sistemas empregam estruturas de árvores de decisão para guiar os usuários através das técnicas de TCC, enfatizando o acompanhamento de sintomas e o gerenciamento do humor. Esse *design* contrasta com modelos gerativos mais complexos, como Coral (100) e o *chatbot* baseado em GPT-2 discutido em (119), que utilizam modelos de aprendizado profundo para gerar respostas empáticas, tornando-os mais adequados para contextos que exigem suporte emocional. A dependência da lógica baseada em regras está alinhada com as descobertas de que componentes estruturados, como a "Lógica Baseada em Regras", estão frequentemente associados aos papéis psicoterapêuticos e de avaliação, oferecendo previsibilidade e consistência essenciais na TCC.

Muitos *chatbots* voltados para propósitos terapêuticos, como o ClientBot (112), que treina conselheiros, e o EDRA (104), que auxilia na gestão da saúde mental, empregam modelos de redes neurais como redes Long Short-Term Memory (LSTM) e redes neurais profundas para classificação de intenções e geração de respostas. Da mesma forma, CARO (42) utiliza múltiplas camadas LSTM para gerar diálogos empáticos, enquanto KokoBot (79) reutiliza dados de suporte entre pares usando técnicas de recuperação de informações, como a Distância do Transportador de Palavras (WMD) com *embeddings* word2vec. Esses modelos mais avançados refletem o alinhamento com papéis que requerem interações dinâmicas, flexíveis e contextuais, ressoando com a posição dos componentes "Componente Gerativo" e "Reconhecimento de Emoção/Sentimento" na Análise de Correspondência.

Chatbots como o CoachAI (29), que visam a saúde física juntamente com o bem-estar mental, utilizam modelos de aprendizado de máquina como Máquinas de Vetores de Suporte (SVMs) para agrupamento de usuários, personalizando ainda mais as intervenções com base em dados comportamentais. Da mesma forma, o *chatbot* descrito em (121) usa SVMs para monitoramento de saúde mental em mulheres perinatais, demonstrando a versatilidade dos modelos SVM em diferentes casos de uso. De modo similar, PRERONA (49) utiliza Processamento de Linguagem Natural (PLN) combinado com algoritmos de correspondência de padrões para suporte em saúde mental em múltiplos idiomas. Essas aplicações ilustram a importância dos componentes "Recuperação de Conhecimento" e "Classificação de Intenções", comumente associados a papéis educativos e de suporte, conforme identificado na análise.

O estudo de bots como Robo (77), projetado para suporte a dependência de opioides, mostra o uso de redes de média profunda para codificação de sentenças e correspondência

de consultas, enquanto MIRA (126) utiliza o modelo Dual Intent and Entity Transformer (DIET) para aprimorar a detecção de intenções e extração de entidades. Esses modelos são otimizados para abordar desafios específicos de saúde mental através de recuperação de informações precisa e suporte, alinhando-se com as descobertas que destacam o papel significativo da classificação de intenções em contextos de avaliação e diagnóstico.

Enquanto *chatbots* como Vincent (63) e Maxx (26) dependem de respostas predefinidas ou estruturas básicas de PLN (e.g., Google Dialogflow), outros modelos como Dejal@bot (81) e Psyche Conversa (107) integram interpretações probabilísticas mais avançadas e modelos de aprendizado profundo para lidar com técnicas de TCC e diagnósticos de saúde mental. Essa variação na complexidade tecnológica ressalta as descobertas da Análise de Correspondência, onde modelos mais simples se alinham mais com intervenções estruturadas e modelos gerativos mais complexos com papéis de suporte dinâmicos.

O *chatbot* PTSDialogue (40), construído para gestão de TEPT, utiliza um sistema de estados finitos, refletindo uma abordagem mais estruturada em comparação com modelos como os de (99), que usam arquiteturas Seq2Seq e Hierarchical Encoder-Decoder (HRED) para gerar interações mais flexíveis e direcionadas ao contexto. Da mesma forma, Vik (15) usa classificação de intenções e reconhecimento de entidades, sendo um modelo mais simples comparado ao PAL (16), que emprega extração de persona e decodificação controlada para suporte emocional personalizado. Essas diferenças ilustram a variabilidade no *design* de chatbots, onde a lógica estruturada muitas vezes se associa a papéis que exigem consistência, e técnicas gerativas avançadas se alinham com necessidades de engajamento emocional personalizado.

Chatbots como o MyUBot e o modelo descrito em (5) integram sistemas baseados em regras com PLN, enquanto sistemas mais sofisticados como MediBot (56) e Mental Ease (36) empregam algoritmos de aprendizado de máquina e análise de sentimento para adaptar as respostas com base nos inputs e estados emocionais dos usuários. Isso reflete a ampla aplicabilidade de modelos híbridos que combinam elementos estruturados e flexíveis, atendendo a diversos papéis em saúde mental, conforme destacado nos testes.

Modelos mais recentes, como MemoryBank em SiliconFriend, incorporam mecanismos de armazenamento e recuperação de memória para aprimorar a capacidade do *chatbot* de lembrar interações passadas, permitindo respostas mais personalizadas e contextualizadas ao longo do tempo. De modo similar, o agente STEF (120) utiliza geradores de múltiplas fontes para produzir respostas de suporte em terapias digitais, demonstrando um foco crescente na fusão emocional e estratégias personalizadas para intervenções em saúde mental. Essas capacidades avançadas se alinham com a tendência observada em direção a papéis de *chatbot* cada vez mais dinâmicos e empáticos, mostrando a evolução das tecnologias de *chatbot* para responder às necessidades complexas da saúde mental.

A principal percepção dessas implementações variadas é que os componentes dos

chatbots não são escolhidos arbitrariamente, mas estrategicamente alinhados com os papéis específicos que esses sistemas se propõem a cumprir. As escolhas técnicas, desde sistemas baseados em regras até redes neurais profundas, refletem as diversas necessidades nas intervenções em saúde mental, reforçando a importância de uma abordagem personalizada no *design* de *chatbots* que integre as tecnologias mais adequadas para a aplicação-alvo. Conforme a Análise de Correspondência e os testes de resíduos sugerem, os modelos de *chatbot* mais eficazes harmonizam suas bases técnicas com os requisitos detalhados dos papéis de saúde mental que pretendem apoiar.

2.4 RQ2: QUÃO ROBUSTAS SÃO AS EVIDÊNCIAS QUE SUSTENTAM A EFICÁCIA DESSES *CHATBOTS* E QUAIS ASPECTOS ESPECÍFICOS ESTÃO SENDO AVALIADOS?

Iniciamos esta seção apresentando uma visão geral dos dados na Tabela 22, que mostra a distribuição de frequência dos diferentes tipos de experimentos em relação aos Resultados de Interesse. Para uma análise mais aprofundada, utilizamos a Tabela 23, que apresenta os resíduos padronizados do qui-quadrado. Esses resíduos são essenciais para avaliar a força e a direção das associações entre tipos de experimentos e resultados, indicando quais relações são estatisticamente significativas. Para controlar o risco de inflacionar a taxa de erro tipo I devido ao grande número de comparações, aplicamos a correção de Bonferroni ao nível de significância, ajustando o critério de decisão para identificar associações relevantes.

Algumas associações se destacam ao examinar os resíduos padronizados, devido à sua significância estatística e relevância contextual. Em primeiro lugar, há uma associação negativa significativa entre Resultados de Saúde Mental/ Comportamentais e Avaliações Baseadas em Modelo e Métricas Automatizadas, com um resíduo de -2.3551. Esse valor sugere que os resultados de saúde mental e comportamentais estão sub-representados em estudos que utilizam métricas automatizadas e baseadas em modelo. Esse achado pode indicar uma limitação metodológica, pois essas abordagens automatizadas podem não capturar adequadamente a complexidade das variáveis comportamentais e de saúde mental, refletindo uma possível desconexão entre a metodologia e a natureza dos fenômenos analisados.

A relação entre Desempenho do Sistema (Intrínseco) e Pesquisa Experimental revela um resíduo de -2.7547, indicando que a pesquisa experimental é aplicada com menos frequência do que o esperado para a avaliação do desempenho intrínseco do sistema. Essa ausência significativa de associação sugere que o controle experimental pode ser desafiador ou menos apropriado ao lidar com variáveis intimamente integradas ao funcionamento interno do sistema, onde métodos modelados e automatizados, como indicado pelo resíduo positivo significativo de 5.1968, parecem mais prevalentes. Esse contraste destaca a

Tabela 22 – Tabela de Contingência dos Tipos de Experimentos e Resultados de Interesse (ER = Pesquisa Experimental; MME = Avaliações de Métodos Mistos; MAE = Avaliações Baseadas em Modelo e Métricas Automatizadas; ND = Descrição Não Avaliativa; OR = Pesquisa Observacional; UCE= Avaliações Focadas no Usuário.)

Resultados de Interesse	**ER**	**MME**	**MAE**	**ND**	**OR**	**UCE**
Resultados de Saúde Mental/ Comportamentais	11	3	3	4	4	10
Desempenho do Sistema (Extrínseco)	6	2	13	3	0	12
Desempenho do Sistema (Intrínseco)	0	2	32	0	0	9
Confiança e Construção de Relacionamento	5	0	0	0	1	5
Engajamento e Adesão do Usuário	4	1	0	2	2	2

Fonte: Elaborado pelo autor (2024).

importância de adaptar a metodologia ao tipo de resultado analisado, reforçando que métricas automatizadas podem fornecer uma avaliação mais precisa e eficiente de variáveis intrínsecas ao sistema.

Por outro lado, observamos uma associação positiva significativa entre Confiança e Construção de Relacionamento e Pesquisa Experimental, com um resíduo de 2.1954, indicando uma preferência notável pelo uso de metodologias experimentais para investigar confiança e construção de relacionamento. Esse resultado sugere que a natureza controlada e manipulativa da pesquisa experimental é considerada valiosa para explorar interações complexas e variáveis comportamentais em contextos controlados, capturando nuances que outras abordagens podem não revelar com a mesma clareza.

Um padrão distinto é observado entre Engajamento e Adesão do Usuário e Pesquisa Observacional, onde o resíduo positivo de 1.8577 sugere que abordagens observacionais são preferidas para estudar engajamento e adesão dos usuários. Essa escolha metodológica parece alinhada com a necessidade de capturar o comportamento em ambientes naturais, proporcionando *insights* sobre como os usuários interagem com sistemas e produtos sem a intervenção experimental direta que poderia alterar seus comportamentos naturais.

A análise também destaca uma associação positiva significativa entre Satisfação e

Tabela 23 – Tabela de Resíduos Padronizados para Tipos de Experimentos e Resultados de Interesse (ER = Pesquisa Experimental; MME = Avaliações de Métodos Mistos; MAE = Avaliações Baseadas em Modelo e Métricas Automatizadas; ND = Descrição Não Avaliativa; OR = Pesquisa Observacional; UCE= Avaliações Focadas no Usuário.)

Resultados de Interesse	**ER**	**MME**	**MAE**	**ND**	**OR**	**UCE**
Resultados de Saúde Mental/ Comportamentais	1.9409	0.4886	-2.3551	1.3529	1.5773	-0.5617
Desempenho do Sistema (Extrínseco)	-0.1400	-0.2158	0.5992	0.6063	-1.3805	-0.0806
Desempenho do Sistema (Intrínseco)	-2.7547	-0.4690	5.1968	-1.5904	-1.5088	-1.4805
Confiança e Construção de Relacionamento	2.1954	-0.8437	-1.8343	-0.8044	0.5473	0.6437
Engajamento e Adesão do Usuário	1.4777	0.3416	-1.8343	1.6819	1.8577	-0.9049
Satisfação e Feedback do Usuário	-0.8165	0.5394	-1.9846	-0.7071	0.1491	2.4663

Fonte: Elaborado pelo autor (2024).

Feedback do Usuário e Avaliações Focadas no Usuário, com um resíduo de 2.4663. Isso evidencia um forte alinhamento metodológico, onde avaliações focadas no usuário são empregadas para explorar diretamente a satisfação e o feedback, refletindo a relevância dessas abordagens para capturar percepções subjetivas e detalhadas dos usuários em relação à sua experiência com os sistemas. Esse alinhamento é crucial, pois assegura que as metodologias empregadas sejam sensíveis às dimensões subjetivas que determinam a aceitação e usabilidade do sistema.

Além das associações significativas, algumas relações mostraram resíduos próximos de zero, sugerindo uma relação neutra ou falta de associação estatisticamente relevante. Por exemplo, Desempenho do Sistema (Extrínseco) e Pesquisa Experimental apresentaram um resíduo de -0.1400, indicando que a frequência observada está muito próxima do esperado sob a independência, sugerindo que essa combinação de tipo de experimento e resultado não apresenta uma associação forte o suficiente para se destacar na análise.

De modo geral, os resultados indicam uma clara tendência de adaptação metodológica, onde as escolhas de tipos de experimentos estão bem alinhadas com as características

dos resultados avaliados. Isso reflete um entendimento técnico por parte dos pesquisadores sobre a importância de selecionar abordagens que melhor correspondam às necessidades dos diferentes tipos de resultados. No entanto, a análise também aponta para áreas que poderiam se beneficiar de maior diversificação metodológica, como a sub-representação de métodos automatizados na avaliação da saúde mental e comportamental, indicando uma potencial oportunidade para expansão e inovação metodológica.

Em resumo, a análise dos resíduos padronizados do qui-quadrado permite a identificação de associações significativas e uma compreensão de como as metodologias são escolhidas de acordo com as características dos resultados, destacando tanto adequações quanto lacunas potenciais que podem orientar investigações futuras. Essa abordagem crítica enfatiza a importância de uma escolha metodológica bem fundamentada para capturar robustamente as nuances dos fenômenos em estudo.

2.4.1 Robustez das evidências que sustentam a eficácia dos *chatbots* em saúde mental

Para avaliar a robustez das evidências que sustentam a eficácia dos *chatbots* em saúde mental, analisamos um conjunto de ensaios clínicos randomizados (ECRs) conduzidos para avaliar a efetividade de diversos *chatbots* em intervenções de saúde mental. As evidências coletadas nesses estudos apontam para vários achados significativos sobre o potencial e as limitações das intervenções baseadas em chatbots, embora o nível geral de evidência varie entre os estudos.

O ensaio de (30) avaliou o Woebot, um *chatbot* projetado para aliviar sintomas de depressão e ansiedade. O ECR comparou os efeitos do Woebot a um grupo controle com informações apenas. Os resultados mostraram reduções significativas nos sintomas de depressão, medidos pelo PHQ-9, para participantes que interagiram com o Woebot. Embora os sintomas de ansiedade tenham melhorado em ambos os grupos, o estudo não encontrou uma diferença significativa entre eles, sugerindo que o Woebot pode ser particularmente eficaz no tratamento dos sintomas depressivos, mas são necessárias mais pesquisas para esclarecer seu impacto sobre a ansiedade.

Em um ECR piloto conduzido por (44), o *chatbot* SELMA foi avaliado quanto ao seu impacto na limitação causada pela dor e em desfechos secundários, como bem-estar e intensidade da dor. Embora o desfecho primário não tenha mostrado diferença significativa entre os grupos de intervenção e controle, houve uma correlação positiva entre a intenção dos participantes de mudar o comportamento e a redução da limitação e intensidade da dor. O estudo indicou altos níveis de aceitação do chatbot, embora alguns participantes tenham expressado insatisfação com a falta de opções de entrada de texto livre, destacando limitações de interação com o usuário.

Em um estudo mais extenso realizado por (81), o Dejal@bot foi testado em um ECR envolvendo mais de 500 participantes para avaliar sua eficácia na promoção da cessação do

tabagismo. O desfecho primário, a abstinência contínua de fumar em seis meses, mostrou uma taxa mais alta no grupo de intervenção em comparação ao grupo controle (26% vs. 18.8%). No entanto, os resultados foram marginalmente significativos (P=0.05). O estudo também relatou melhorias na qualidade de vida, mas essas mudanças não foram estatisticamente significativas, sugerindo que o Dejal@bot pode ajudar na cessação do tabagismo até certo ponto, mas a robustez dessa evidência permanece incerta devido à significância limítrofe dos achados.

O protocolo de estudo de (103) descreve a avaliação do ELME, um *chatbot* baseado em regras projetado para reduzir o estresse por meio de mindfulness e psicoeducação. Como este estudo ainda está em andamento, não há resultados disponíveis no momento. O *design* visa medir a redução do estresse como desfecho primário, com desfechos secundários incluindo ansiedade, depressão e bem-estar. O protocolo demonstra uma abordagem experimental rigorosa, mas conclusões sobre a eficácia do *chatbot* só poderão ser tiradas após a publicação dos resultados do estudo.

(97) avaliou o *chatbot* Emohaa em um ensaio que mediu depressão, ansiedade, afeto negativo e insônia. O estudo encontrou melhorias significativas na depressão, afeto negativo e insônia no grupo de intervenção em comparação ao grupo controle. No entanto, as reduções na ansiedade não foram significativamente diferentes entre os grupos na análise por intenção de tratar (ITT). Esses resultados fornecem evidências moderadas de que o Emohaa é eficaz no alívio de sintomas específicos de saúde mental, particularmente depressão e insônia, embora seu impacto sobre a ansiedade pareça menos conclusivo.

O ECR conduzido por (114) comparou o ChemoFreeBot à educação liderada por enfermeiros e ao atendimento rotineiro no suporte a pacientes em quimioterapia para autocuidado e manejo de efeitos colaterais. As mulheres no grupo ChemoFreeBot relataram significativamente menos e menos graves sintomas, tanto físicos quanto psicológicos, em comparação aos grupos liderados por enfermeiros ou cuidados de rotina. O estudo apresenta fortes evidências de que o ChemoFreeBot pode reduzir efetivamente o desconforto e a gravidade dos efeitos colaterais da quimioterapia, com melhorias notáveis nos comportamentos de autocuidado, apoiando a utilidade do *chatbot* em um contexto clínico.

(125) conduziu um ECR para avaliar o impacto de um *chatbot* na taxa de conclusão de um programa de terapia cognitivo-comportamental baseado na internet (iCBT). O grupo de intervenção, que utilizou o chatbot, teve taxas de conclusão significativamente maiores do que o grupo controle (34.8% vs. 19.2%). No entanto, não houve diferenças significativas nas pontuações de depressão ou ansiedade entre os grupos. Embora o *chatbot* melhore a adesão ao programa, seu efeito sobre os desfechos psicológicos permanece incerto, sugerindo que uma exploração adicional de seu valor terapêutico é necessária.

Por fim, (115) avaliou duas interfaces de *chatbot* — uma apenas de texto e outra com avatar digital humano — usando medidas de usabilidade e resposta emocional. Embora o

chatbot apenas de texto tenha sido classificado como mais amigável, com uma pontuação mais alta na Escala de Usabilidade do Sistema (SUS-10) (75.34 vs. 64.80), não houve diferenças significativas no engajamento emocional entre as interfaces. Este estudo destaca a importância da usabilidade no *design* de chatbots, mas não fornece evidências fortes de benefícios emocionais ou psicológicos.

De forma geral, os resultados desses experimentos sugerem que, embora os *chatbots* de saúde mental tenham o potencial de impactar positivamente os desfechos de saúde mental, a força e a consistência das evidências variam entre diferentes contextos e intervenções específicas.

O Woebot mostrou reduzir significativamente os sintomas de depressão, mas seu efeito sobre a ansiedade foi menos claro (30). Da mesma forma, o SELMA demonstrou eficácia limitada na redução de limitações causadas pela dor, embora tenha mostrado uma relação positiva entre as intenções comportamentais dos participantes e a redução da intensidade da dor (44). O Dejal@bot demonstrou um aumento marginalmente significativo nas taxas de cessação do tabagismo, mas a robustez dessa evidência é questionável devido à significância limítrofe (P=0.05) (81). O *chatbot* Emohaa provou ser eficaz na redução de depressão, insônia e afeto negativo, embora seu impacto na ansiedade não tenha sido significativo na análise ITT (97). O ChemoFreeBot mostrou fortes evidências de eficácia na redução da frequência, gravidade e desconforto dos efeitos colaterais da quimioterapia, superando tanto a educação liderada por enfermeiros quanto o atendimento de rotina (114). O *chatbot* estudado por (125) melhorou a adesão ao programa de iCBT, mas não mostrou efeitos significativos sobre a depressão ou ansiedade. Finalmente, o estudo de (115) encontrou que, embora a interface do *chatbot* apenas de texto fosse mais amigável, não foram observadas diferenças significativas nas respostas emocionais entre as interfaces de chatbot.

Em conclusão, as evidências que sustentam a eficácia dos *chatbots* de saúde mental são promissoras, mas inconsistentes. Alguns chatbots, como o Woebot e o ChemoFreeBot, mostraram evidências fortes de sua eficácia em contextos específicos, como gerenciamento de depressão e redução de efeitos colaterais da quimioterapia. Outros, como o SELMA e o *chatbot* de iCBT, demonstraram efeitos limitados ou inconclusivos sobre os principais desfechos. Mais pesquisas são necessárias para determinar o impacto a longo prazo dessas intervenções, especialmente em termos de sua escalabilidade, eficácia no mundo real e aplicabilidade em populações e condições de saúde mental diversas. Além disso, muitos dos estudos destacaram a importância da usabilidade e da satisfação do usuário, sugerindo que o sucesso dessas intervenções pode depender tanto do engajamento do usuário quanto dos mecanismos terapêuticos subjacentes.

2.5 CONSIDERAÇÕES FINAIS

A primeira questão de pesquisa (RQ1) explorou como os modelos de *chatbots* variam em implementação em diferentes casos de uso. Os resultados demonstram que o *design* e a funcionalidade dos modelos de *chatbots* estão intrinsecamente ligados aos seus papéis pretendidos e aos resultados específicos que visam alcançar. A análise das relações entre os componentes do modelo e os alvos de saúde mental indica que nenhum componente de modelo de *chatbot* é universalmente eficaz em todos os contextos. A Recuperação de Conhecimento, por exemplo, mostrou uma associação significativa com o Bem-estar Psicológico, enquanto a Lógica Baseada em Regras esteve moderadamente ligada ao gerenciamento de Distúrbios Gerais. Esses resultados sugerem que diferentes alvos de saúde mental requerem abordagens tecnológicas distintas e que o uso de um modelo híbrido — que combina múltiplos componentes — pode fornecer a solução mais eficaz para abordar uma gama mais ampla de questões de saúde mental.

A Análise de Correspondência destacou ainda mais a complexidade dessas relações ao revelar padrões de associações entre os componentes do modelo e os alvos de saúde mental. Os resultados indicaram que, embora alguns componentes, como a Classificação de Intenções e o Reconhecimento de Emoção/Sentimento, sejam amplamente aplicados em diferentes contextos, sua generalização excessiva pode limitar sua eficácia em intervenções de saúde mental específicas. Por exemplo, a Classificação de Intenções mostrou algum alinhamento com intervenções relacionadas ao humor, mas sua ampla aplicabilidade não aprimorou significativamente nenhum resultado de saúde mental específico. A importância da Recuperação de Conhecimento na promoção do Bem-estar Psicológico, confirmada tanto pelos resíduos padronizados quanto pela Análise de Correspondência, destaca o potencial desse componente em casos de uso focados em educação em saúde mental e promoção do bem-estar a longo prazo.

Um ponto chave desta pesquisa é a necessidade de modelos de *chatbots* personalizados que se adaptem às necessidades emocionais e psicológicas únicas dos usuários. Embora alguns componentes demonstrem utilidade em contextos específicos de saúde mental, seu impacto é maximizado quando são combinados em um sistema flexível e responsivo que pode atender às complexidades das questões individuais de saúde mental. Isso tem implicações amplas para a pesquisa e desenvolvimento futuros, sugerindo que os sistemas de *chatbot* não devem ser projetados de forma "tamanho único", mas sim evoluir para modelos personalizados e adaptativos que integrem vários componentes para responder às particularidades de diferentes condições de saúde mental.

Com relação aos papéis desempenhados pelos *chatbots* em intervenções de saúde mental, os resultados também revelam uma associação significativa entre os componentes técnicos dos *chatbots* e suas aplicações práticas. Por exemplo, o componente de Reconhecimento de Emoção/Sentimento foi altamente associado a Ferramentas de Avaliação e

Diagnóstico, o que se alinha com sua relevância para interpretar os estados emocionais dos usuários para diagnósticos mais precisos e contextualizados. Da mesma forma, o Componente Gerativo mostrou uma associação positiva com Técnicas de Comunicação e Aconselhamento, onde a flexibilidade nas respostas é fundamental. A capacidade dos sistemas de *chatbot* de gerar respostas empáticas e contextualmente apropriadas é crítica nesses papéis, e essa associação indica que o desenvolvimento futuro de *chatbots* nessas áreas pode se beneficiar ao aprimorar ainda mais as capacidades gerativas.

O papel da Lógica Baseada em Regras em abordagens psicoterapêuticas também merece atenção. Sua natureza estruturada e baseada em regras a torna adequada para intervenções de Terapia Cognitivo-Comportamental (TCC) que requerem interações consistentes e previsíveis. No entanto, os resultados sugerem que esses sistemas baseados em regras podem carecer da flexibilidade necessária para outros papéis, como aqueles que exigem intervenções mais dinâmicas e personalizadas. Isso destaca uma consideração importante no *design* de chatbots: embora sistemas rígidos baseados em regras sejam eficazes em abordagens terapêuticas estruturadas, eles podem precisar ser complementados com tecnologias mais adaptativas, como modelos baseados em aprendizado de máquina, para melhorar sua versatilidade em contextos mais amplos de saúde mental.

A segunda questão de pesquisa (RQ2) abordou a robustez das evidências que sustentam a eficácia desses chatbots, e a análise gerou resultados mistos. Embora haja fortes evidências que sustentam a eficácia de certos chatbots, como Woebot (30) e Chemo-FreeBot (114), especialmente na redução de sintomas de depressão e no gerenciamento de efeitos colaterais da quimioterapia, outros *chatbots* demonstraram efeitos limitados ou inconsistentes. Por exemplo, o *chatbot* SELMA (44) mostrou resultados positivos em alguns desfechos secundários, mas não teve um impacto significativo nos desfechos primários, indicando que, embora os *chatbots* possam ter potencial, sua eficácia pode variar dependendo do contexto e dos resultados específicos de saúde mental que estão sendo direcionados.

Além disso, a análise revelou que as abordagens metodológicas empregadas nesses estudos geralmente estão bem alinhadas com os resultados de interesse. A Pesquisa Experimental foi o método mais robusto para avaliar confiança e construção de relacionamento, enquanto as Avaliações Focadas no Usuário foram mais frequentemente empregadas para avaliar satisfação e *feedback* dos usuários. No entanto, há uma sub-representação notável de metodologias específicas, como Avaliações Baseadas em Modelo e Métricas Automatizadas, na avaliação de resultados de saúde mental e comportamentais. Essa lacuna sugere que pesquisas futuras poderiam se beneficiar da incorporação de abordagens mais automatizadas para avaliar resultados complexos de saúde mental, desde que esses métodos possam ser adaptados para capturar nuances do estado psicológico.

Embora as evidências que sustentam a eficácia dos *chatbots* de saúde mental

sejam promissoras, há áreas claras onde mais pesquisas são necessárias. Os resultados inconsistentes entre diferentes *chatbots* destacam a necessidade de mais ensaios clínicos randomizados (ECRs) rigorosos e em grande escala para fornecer evidências mais robustas sobre a eficácia a longo prazo e a escalabilidade dessas intervenções. Além disso, muitos dos estudos revisados enfatizaram a importância da usabilidade e do engajamento do usuário na determinação do sucesso das intervenções com chatbots. Isso sugere que, mesmo os *chatbots* tecnicamente mais sofisticados falharão em alcançar seu pleno potencial, a menos que sejam projetados com foco na experiência do usuário, o que inclui garantir interfaces intuitivas e respostas empáticas que atendam às necessidades emocionais dos usuários.

Em conclusão, esta pesquisa destaca a variabilidade nas implementações de *chatbots* em diferentes casos de uso de saúde mental e ressalta a importância de alinhar os componentes dos modelos de *chatbot* com resultados específicos de saúde mental. Embora encorajadoras, as evidências que sustentam a eficácia desses *chatbots* ainda estão em desenvolvimento, com várias áreas que requerem investigação adicional. Pesquisas futuras devem focar no refinamento dos modelos de *chatbot* para garantir que sejam tecnicamente sólidos e centrados no usuário, bem como explorar a eficácia a longo prazo e a escalabilidade dessas intervenções no mundo real. Com a combinação apropriada de inovação técnica e *design* focado no usuário, os *chatbots* têm grande potencial para aprimorar significativamente o cuidado em saúde mental, fornecendo suporte acessível, personalizado e eficaz a populações diversas.

3 PROPOSTA INICIAL DE INTERVENÇÃO AUTOGUIADA

Neste capítulo descrevemos um estudo de caso, utilizando dados abertos do contexto de entrevistas motivacionais. Em um primeiro momento, elencamos a literatura relevante no que tange a modelagem de intervenções psicoterapêuticas com a utilização de componentes generativos. Em sequência, elencamos a literatura disponível a respeito de modelos aplicados no contexto de entrevista motivacional, com o objetivo de fundamentar e guiar nossa abordagem. Depois, descrevemos o processo metodológico de nossa abordagem, avaliamos os resultados e traçamos considerações.

3.1 FUNDAMENTAÇÃO NA LITERATURA

A modelagem de conversações com padrões e protocolos bem definidos apresenta diversas dificuldades, especialmente quando se trata de abordagens psicoterapêuticas que envolvem componentes generativos. Conforme verificado na revisão sistemática realizada (15), apenas 11 estudos relatam modelos que utilizam um componente generativo para desempenhar um papel em abordagens psicoterapêuticas. Na Tabela 24, analisamos esses estudos individualmente e identificamos que 6 dos 11 modelos utilizam componentes adicionais para alcançar seus objetivos.

Um exemplo é o modelo apresentado por (107) que inclui módulos de coleta de dados, como um *keylogger*, um módulo de chat para interação com o usuário e um módulo de detecção de doenças mentais baseado em aprendizado profundo. Os modelos de aprendizado profundo utilizados incluem CNN-LSTM, BiLSTM e BERT, com mecanismos de atenção para melhorar a precisão da classificação. O propósito do Psyche é auxiliar indivíduos a identificar suas condições de saúde mental e oferecer terapia apropriada por meio de um *chatbot* baseado em aprendizado profundo, além de monitorar atividades em redes sociais para detectar sinais precoces de doenças mentais. O *chatbot* foi treinado com o Reddit Mental Health Dataset, que contém postagens de 28 subreddits, incluindo 15 dos maiores grupos de apoio ao bem-estar mental do mundo. O desempenho do *chatbot* foi avaliado usando métricas como acurácia, precisão, recall, F1-score e área sob a curva ROC (AUC) sendo que o classificador baseado em BERT apresentou o melhor desempenho com uma acurácia de 75.3% e um F1-score de 0.71.

Outro modelo relevante é o descrito em (11), que utiliza o modelo Serena, um Transformer Seq2Seq com 2.7 bilhões de parâmetros. Este modelo inclui algoritmos de pós-processamento que utilizam modelos menores baseados em Transformer para detectar contradições, reconhecer linguagem tóxica e evitar respostas repetitivas. Serena emprega a técnica de *beam search* para selecionar a resposta mais adequada entre uma lista de candidatas. O objetivo do modelo é oferecer aconselhamento psicológico em uma plataforma virtual, proporcionando uma alternativa acessível e de baixo custo à terapia tradicional.

Serena foi treinada com o Pushshift Reddit Dataset, que inclui 651 milhões de submissões e 5.6 bilhões de comentários, e posteriormente refinada com 14.300 pares de prompts de pacientes e respostas de conselheiros extraídas de transcrições de aconselhamento e psicoterapia. A avaliação do modelo foi realizada por meio de pesquisas com usuários, que avaliaram a compreensão das mensagens pelo modelo e a utilidade das respostas, além de testes internos com um banco de dados crescente de prompts e respostas.

No caso de (97), o Emohaa utiliza um modelo de diálogo em larga escala, denominado ES-Bot para gerar respostas com base na entrada do usuário e na estratégia de suporte emocional escolhida. O Emohaa, que também utiliza o CBT-Bot, combina um sistema baseado em templates com exercícios de Terapia Cognitivo-Comportamental (TCC) para fornecer suporte cognitivo e emocional aos usuários. O objetivo é reduzir sintomas de sofrimento mental, incluindo depressão, ansiedade, afeto negativo e insônia, por meio de exercícios de TCC e conversas de apoio emocional. O ES-Bot foi treinado usando o ESConv dataset, contendo conversas de suporte emocional baseadas na Helping Skills Theory. A avaliação foi realizada por meio de um ensaio clínico randomizado com 247 participantes, medindo mudanças em depressão, ansiedade, afeto negativo e insônia utilizando questionários padronizados (PHQ-9, GAD-7, PANAS, ISI).

O modelo desenvolvido em (36) utiliza um modelo de aprendizado profundo baseado na arquitetura Transformer, combinado com técnicas de Processamento de Linguagem Natural, como tokenização, análise de sentimento e classificação de texto. O *chatbot* também inclui um sistema de gerenciamento de diálogo baseado em regras e visa oferecer suporte conversacional a indivíduos com problemas de saúde mental, particularmente ansiedade leve e depressão. O *chatbot* foi treinado com os datasets Cornell Movie Dialog Corpus e OpenSubtitles Corpus, consistindo em diálogos extraídos de filmes. A eficácia do *chatbot* foi avaliada utilizando métricas como precisão, recall e F1-score, além de um estudo com usuários para avaliar a efetividade do suporte fornecido.

Já ClientBot (112) utiliza dois modelos de rede neural LSTM: um modelo seq2seq treinado em transcrições de filmes e transcrições de psicoterapia, e um LSTM mais simples treinado apenas em transcrições de psicoterapia. O ClientBot simula respostas de pacientes e fornece *feedback* em tempo real sobre o uso de habilidades de aconselhamento por terapeutas em treinamento. O treinamento foi realizado com 2354 transcrições de psicoterapia e transcrições de filmes em inglês (open-subs dataset). A avaliação do modelo ocorreu por meio de um ensaio clínico randomizado com 151 não-terapeutas, medindo a eficácia do treinamento pela utilização de perguntas abertas e reflexões pelos participantes.

Por fim, o *chatbot* KEMI (23) inclui um módulo de aquisição de conhecimento que recupera informações relevantes de um grafo de conhecimento em saúde mental (HEAL) e um módulo de geração de respostas que utiliza uma abordagem seq2seq para criar respostas de iniciativa mista. O objetivo do KEMI é fornecer suporte emocional

e ajudar os usuários a explorar e enfrentar seus problemas. O *chatbot* foi treinado com os datasets EMPATHETICDIALOGUES e ESConv, que contêm diálogos de suporte emocional. A avaliação incluiu métricas automáticas como Perplexity, BLEU e ROUGE, além de avaliações humanas focadas na fluência, informatividade e apoio das respostas geradas pelo modelo.

Tabela 24 – Modelos com componente Generativa que desempenham alguma abordagem psicoterapêutica

Estudo	**Ano**	**Intent Classification Component**	**Emotion/Sentiment Recognition Component**	**Knowledge Retrieval Component**	**Logic of Rules Component**
(107)	2022	Sim	Sim		
(99)	2021				
(129)	2022				
(11)	2023		Sim		
(119)	2021				
(97)	2023	Sim	Sim		Sim
(50)	2023				
(36)	2023		Sim		
(38)	2023				
(112)	2019	Sim			Sim
(23)	2023			Sim	

Fonte: Elaborado pelo autor (2024).

Agora, voltando aos trabalhos que lidam com modelos puramente generativos, temos o seguinte panorama. Em (99), pe apresentado um *chatbot* que utiliza modelos generativos especificamente as arquiteturas Sequence-to-Sequence (Seq2Seq) e Hierarchical Encoder-Decoder (HRED) para gerar respostas. Esses modelos são treinados utilizando uma combinação de aprendizado supervisionado e aprendizado por reforço com o objetivo de otimizar recompensas conversacionais de longo prazo. O dataset MotiVAte foi criado a partir da raspagem e modificação de conversas do fórum online PsychCentral que contém discussões relacionadas à saúde mental, especialmente depressão. O conjunto de dados é composto por conversas diádicas entre usuários e o assistente virtual. A performance do *chatbot* foi avaliada por meio de métricas automáticas, como BLEU score, perplexidade e métricas baseadas em *embeddings*, além de uma avaliação humana. Os avaliadores humanos classificaram a qualidade das respostas geradas com base na fluência, adaptabilidade, capacidade de lidar com ambiguidades, criatividade e no progresso das conversas.

Em (129), o modelo utiliza componentes que incluem um modelo de linguagem neural GPT-2 ajustado para gerar respostas, junto com um sistema Whole Dialogue History (WDH) que garante a coerência de longo prazo ao analisar todo o histórico de diálogo.

O sistema WDH também oferece explicabilidade ao visualizar o processo de tomada de decisão. O objetivo do *chatbot* MotiVAte é atuar como um assistente virtual servindo como o primeiro ponto de contato para usuários que enfrentam depressão ou desânimo, proporcionando respostas motivacionais e afirmativas e criando um ambiente seguro para os usuários compartilharem seus pensamentos e buscarem ajuda de forma anônima. O *chatbot* foi treinado em um corpus multilíngue de diálogos de coaching coletados por meio de uma plataforma Wizard of Oz, Incluindo diálogos em espanhol, francês, norueguês e inglês, com traduções entre esses idiomas. O desempenho do *chatbot* foi avaliado tanto por métricas automáticas quanto por avaliações humanas. Experimentos de interação com especialistas em coaching foram conduzidos para avaliar a usabilidade do sistema e seu impacto emocional nos usuários.

Já o estudo (119) apresenta um *chatbot* baseado no modelo GPT-2 da OpenAI, um modelo generativo pré-treinado de forma não supervisionada, que foi ajustado com dados específicos de domínio para melhorar seu desempenho em contextos terapêuticos. O *chatbot* é projetado para realizar sessões de coaching motivacional, ajudando os usuários a refletirem sobre seus objetivos, obstáculos e ações potenciais para alcançar mudanças comportamentais. O modelo busca replicar as estratégias de coaching de profissionais de forma totalmente orientada por dados. Para o ajuste fino foram usadas 306 transcrições de sessões de terapia entre cuidadores familiares de indivíduos com demência e terapeutas realizando Terapia de Solução de Problemas. A avaliação do *chatbot* baseou-se em três medidas meta-informacionais: a proporção de saídas sem palavras, o comprimento das respostas e os componentes de sentimento. O modelo ajustado foi comparado com o modelo pré-treinado e com as respostas originais dos terapeutas.

O trabalho (50) discute o uso do ChatGPT no contexto de suporte à saúde mental de cuidadores envolvendo-os em conversas terapêuticas especificamente usando técnicas da Terapia de Solução de Problemas. Embora o artigo não forneça detalhes específicos sobre a avaliação do desempenho do ChatGPT no contexto de saúde mental menciona que ferramentas baseadas em IA como o ChatGPT demonstraram potencial na redução da depressão em outros estudos como no ensaio do Woebot.

Por fim, o estudo (38) apresenta o modelo Friendy, que é uma rede neural profunda com uma topologia de camadas densas, composta por 5 camadas, incluindo 3 camadas ocultas com funções de ativação ReLU e uma camada de saída com função softmax. Para evitar overfitting, foram adicionados nós de dropout. Friendy é projetado para atuar como um companheiro virtual de terapia para ajudar crianças autistas a se sentirem mais confortáveis e cooperativas durante as sessões de terapia. O modelo visa melhorar a experiência terapêutica engajando as crianças em conversas interativas personalizadas de acordo com seus interesses e preocupações. O conjunto de dados utilizado para treinar o Friendy foi criado sob medida, contendo 150 contextos com pelo menos 2 padrões cada, desenvolvido com a contribuição de psiquiatras infantis e outros profissionais, focando em

contextos comuns e padrões de conversa observados em sessões terapêuticas com crianças autistas. O desempenho do *chatbot* foi avaliado por meio de testes de acurácia em um conjunto de dados de treino e validação, alcançando uma acurácia de 97% no treinamento e 80.5% na validação. Além disso, a estrutura foi testada em sessões reais de terapia com crianças autistas, e o *feedback* foi coletado de profissionais envolvidos.

Vários estudos exploraram a aplicação de agentes conversacionais para fornecer intervenções de EM. Por exemplo, agentes conversacionais foram desenvolvidos em contextos médicos e de reabilitação. Em (69), foi criado um agente sensível ao contexto para apoiar a equipe médica e indivíduos com deficiências de movimento em reabilitação, utilizando técnicas de EM para melhorar o engajamento do paciente e a adesão ao protocolo. De maneira semelhante, em (13) introduz-se LvL UP 1.0, um agente baseado em smartphone que fornece intervenções holísticas de estilo de vida para prevenir doenças não transmissíveis e distúrbios mentais comuns, empregando EM para promover escolhas de estilo de vida mais saudáveis.

No apoio à saúde mental, é apresentado em (99) um sistema de diálogo motivacional para indivíduos com Transtorno Depressivo Maior (MDD), focado em criar um ambiente seguro para autoexpressão e ajuda inicial. Esse sistema utilizou modelos generativos como Seq2Seq e Encoder Decoder Hierárquico (HRED) para gerar respostas contextualmente apropriadas e motivacionais. Complementando isso, em (76) desenvolve-se Help Me Heal, um sistema de diálogo educado e empático reforçado para aconselhamento em saúde mental e legal voltado para vítimas de crime, que mostrou engajamento efetivo e respostas empáticas.

Aplicações educacionais de agentes de EM também têm sido significativas. O estudo (113) desenvolve o ClientBot, um agente conversacional similar a um paciente, projetado para treinar participantes não terapeutas em habilidades básicas de aconselhamento por meio de prática interativa, demonstrando melhorias no uso de reflexões e perguntas abertas. Adicionalmente, (88) apresenta um *chatbot* para gerenciamento de estresse entre estudantes de pós-graduação, facilitando auto-reflexão significativa e aprimoramento motivacional.

Intervenções de mudança de comportamento e estilo de vida têm sido outra área chave. Em (4) desenvolve-se um *chatbot* baseado em entrevista motivacional para envolver fumantes em autorreflexão sobre os prós e contras do tabagismo, demonstrando eficácia em iniciar diálogos significativos que promovem a mudança de comportamento. De maneira semelhante, em (111), desenvolve-se MICA, um agente conversacional para pais, com o objetivo de incutir hábitos alimentares saudáveis em seus filhos, facilitando a autorreflexão e motivação entre os pais. Esses estudos demonstram coletivamente a aplicabilidade de agentes conversacionais e seu potencial em fornecer intervenções de EM.

Ao contrário dos trabalhos anteriores, nosso estudo é a primeira aplicação de modelos LLM no contexto de entrevista motivacional em português. Comparamos o

GPT-4 Turbo com modelos de código aberto menores e ajustados para nossa tarefa.

3.2 ARQUITETURA TRANSFORMER

Os Transformers, introduzidos por (118), representam uma inovação significativa no campo do aprendizado profundo. Diferentemente de abordagens anteriores baseadas em redes neurais recorrentes (RNNs) ou convolucionais (CNNs), os Transformers utilizam exclusivamente mecanismos de atenção para capturar dependências dentro de sequências, eliminando a necessidade de processamento sequencial. Essa característica permite que o modelo seja altamente eficiente e escalável, tornando-se adequado para tarefas que envolvem grandes volumes de dados e sequências longas.

A estrutura básica de um Transformer é composta por camadas empilhadas de atenção multi-cabeça e redes *feed-forward* totalmente conectadas. Cada camada é envolvida por conexões residuais e normalização, o que estabiliza o treinamento e melhora a propagação de gradientes. O mecanismo de atenção funciona através da interação de *queries* (Q), *keys* (K) e *values* (V), calculando uma matriz de relevância que pondera as informações mais importantes em uma sequência. A fórmula central que define a atenção escalonada por produto escalar é dada por:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3.1}$$

onde d_k é a dimensão das *keys*, e a normalização por $\sqrt{d_k}$ evita instabilidades durante o treinamento. Para capturar múltiplas relações contextuais simultaneamente, o Transformer utiliza o mecanismo de atenção multi-cabeça, que realiza operações de atenção em diferentes subespaços da representação de dados.

Ao eliminar a dependência de convoluções ou recorrência, os Transformers possibilitam processamento paralelo eficiente, o que revolucionou o campo do processamento de linguagem natural. Essa arquitetura se tornou a base para modelos avançados como o GPT, BERT e LLaMA, destacando-se por sua versatilidade em tarefas como tradução, sumarização e geração de texto. Em resumo, o Transformer combina simplicidade e eficiência, estabelecendo um novo paradigma para a modelagem de sequências em inteligência artificial.

3.3 PROCESSO METODOLÓGICO

Esta seção descreve os materiais e métodos empregados em nosso estudo, detalhando o processo de preparação dos dados, a tradução e formatação do conjunto de dados, e a integração com nossos modelos conversacionais.

3.3.1 Preparação dos Dados

O conjunto de dados utilizado neste estudo é derivado do banco de dados AnnoMI (123), que inclui 133 conversas transcritas de entrevistas motivacionais de alta e baixa qualidade disponíveis em plataformas como YouTube e Vimeo. O banco de dados AnnoMI contém aproximadamente 242.680 tokens e 13.552 enunciados, proporcionando uma base robusta para o treinamento de modelos conversacionais.

Essas transcrições foram realizadas profissionalmente e anotadas por especialistas em entrevista motivacional, detalhando vários atributos ao nível da fala, como interlocutor (terapeuta ou cliente), tópico (o tema da conversa ou transcrição ao qual o enunciado pertence), qualidade da entrevista motivacional (qualidade de EM na transcrição, seja alta ou baixa) e o tipo de fala usada pelo terapeuta (por exemplo, pergunta, informação, reflexão) ou cliente (por exemplo, fala de mudança, neutra, suporte).

Para nosso estudo, as conversas originais do banco de dados AnnoMI foram traduzidas do inglês para o português usando a API do DeepL [1]. Apesar da tradução dos dados pode ser considerada uma limitação do estudo, pois dados traduzidos naturalmente perdem alguns dos significados e relações semânticas originais, essa etapa aborda uma lacuna na disponibilidade de tais recursos para estudantes e profissionais de psicologia de língua portuguesa. Ao tornar nossos modelos acessíveis e relevantes para esse público mais amplo, aumentamos sua aplicabilidade prática tanto em ambientes educacionais quanto clínicos. As traduções foram manualmente checadas, de maneira a avaliar sua qualidade. Só então, a base foi consolidada. Assim, esta também é uma importante contribuição deste trabalho.

Para nosso estudo, as tags distintas presentes no conjunto de dados, como as listadas anteriormente, foram removidas, restando apenas as anotações dos falantes. A ordem dos enunciados foi preservada. Adicionalmente, o tema da conversa não foi usado durante a fase de treinamento, mas essa informação foi mantida para análise posterior, a fim de avaliar a capacidade do modelo de incorporar e modelar esses tópicos.

Após a tradução, os dados foram formatados em JSON, permitindo uma manipulação mais fácil e integração com nossos modelos conversacionais. Essa etapa garante que o conjunto de dados traduzido permaneça consistente com a estrutura e o detalhamento das anotações do banco de dados AnnoMI original, mantendo o rigor e a confiabilidade dos dados para análise e treinamento subsequente dos modelos.

3.4 Modelos

Esta seção descreve Llama 3 8B e Gemma 7B, que são os modelos adotados neste estudo. Detalhamos suas arquiteturas e capacidades.

[1] https://www.deepl.com

3.4.1 Llama 3 8B

Llama 3 é uma coleção de modelos de linguagem fundamentais desenvolvidos pela Meta AI, variando de 8 bilhões a 70 bilhões de parâmetros. O modelo Llama 3 8B é baseado em uma arquitetura de transformador. Suas principais características incluem (54):

- **Arquitetura do Modelo**: O Llama 3 emprega uma arquitetura de transformador somente decodificador com atenção por consulta agrupada (GQA) para aumentar a eficiência de inferência. O modelo utiliza um tokenizador com vocabulário de 128K tokens, o que melhora a eficiência da codificação linguística.
- **Dados de Treinamento**: O modelo é pré-treinado em mais de 15 trilhões de tokens de fontes publicamente disponíveis, incluindo uma porção significativa de dados de alta qualidade em outros idiomas, para apoiar casos de uso multilíngues.
- **Desempenho**: O Llama 3 8B demonstrou forte desempenho em vários benchmarks, superando muitos modelos maiores em cenários reais. Ele é otimizado para benchmarks padrão e aplicações práticas, tornando-o adequado para casos de uso diversos.

3.4.2 Gemma 7B

Gemma é uma família de modelos abertos leves e de última geração desenvolvidos pelo Google, construídos usando pesquisa e tecnologia dos modelos Gemini. O modelo Gemma 7B foi projetado para oferecer alto desempenho, ao mesmo tempo que é acessível para desenvolvedores e dispositivos com restrições de computação. Principais características incluem (73):

- **Arquitetura do Modelo**: Gemma 7B é baseado em uma arquitetura de transformador decodificador com várias melhorias, como Atenção Multi-Consulta, Ativações GeGLU e RMSNorm. Esses aprimoramentos contribuem para a eficiência e o desempenho do modelo.
- **Dados de Treinamento**: O modelo é treinado com um comprimento de contexto de 8.192 tokens, utilizando um conjunto de dados amplo e diversificado para garantir desempenho robusto em várias tarefas.
- **Desempenho**: O Gemma 7B supera modelos de tamanho similar em vários benchmarks, demonstrando fortes capacidades de compreensão de linguagem, raciocínio e segurança. Ele é otimizado para implantação em diversos dispositivos, de laptops a infraestrutura em nuvem, garantindo ampla acessibilidade.

3.4.3 Ajuste Fino Eficiente de Parâmetros (PFET)

Avanços recentes em técnicas de ajuste fino eficiente de parâmetros aumentaram significativamente o desempenho e a eficiência dos grandes modelos de linguagem (LLMs). Duas técnicas-chave empregadas neste trabalho são o RoPE e o LoRA, ambos desempenhando um papel crucial na otimização do ajuste fino do modelo com custo adicional mínimo de parâmetros.

Escalonamento de *embedding* de Posição Rotatória: O RoPE escala *embeddings* posicionais de forma a melhorar as capacidades de extrapolação dos LLMs, particularmente em cenários de treinamento-curto-teste-longo (TSTL). Esse escalonamento melhora a capacidade do modelo de generalizar de sequências de entrada mais curtas para sequências de teste mais longas. Modificando a base rotatória e ajustando o comprimento do contexto, o RoPE aprimora o desempenho do modelo em tarefas como ajuste fino de conversação, onde a extrapolação além dos dados vistos é crítica. No contexto de nossos modelos, o escalonamento RoPE é essencial para garantir que os modelos Gemma e Llama 3 lidem com diálogos longos e de múltiplas etapas típicos da entrevista motivacional (66).

Adaptação de Baixa Ordem: LoRA é uma técnica projetada para reduzir o número de parâmetros treináveis, ao fatorizar as matrizes de atualização de pesos durante o ajuste fino em duas matrizes de baixa ordem. Isso reduz o custo computacional mantendo o desempenho. Neste estudo, o LoRA foi crucial para adaptar eficientemente nossos grandes modelos à tarefa específica de entrevista motivacional sem a necessidade de treinar o modelo inteiro novamente. É particularmente eficaz quando aplicado a mecanismos de atenção, que são essenciais em modelos de diálogo (47).

3.5 EXPERIMENTAÇÃO

O processo de ajuste fino para os modelos Gemma e Llama 3 foi realizado usando o escalonamento RoPE e o LoRA. Utilizamos o conjunto de dados AnnoMI, selecionando apenas os dados conversacionais de maior qualidade para o treinamento, garantindo que as saídas dos modelos fossem o mais precisas e relevantes possível. A base AnnoMI original já possuí a classificação do nível de qualidade da entrevista movacional. Abaixo, fornecemos uma explicação detalhada dos parâmetros utilizados em nossos experimentos, abordando seus impactos individuais no treinamento e desempenho.

3.5.1 Configuração do Treinamento

Os modelos foram ajustados usando uma GPU NVIDIA A100 com 40GB de memória, hospedada no ambiente Google Cloud. Essa configuração de hardware forneceu os recursos computacionais necessários para lidar com a natureza intensiva em memória do processo de ajuste fino para LLMs, especialmente ao incorporar técnicas como checkpointing

de gradiente.

3.5.2 Hiperparâmetros

Nesta seção, apresentamos os principais hiperparâmetros utilizados no treinamento e ajuste fino do modelo, juntamente com as motivações para suas escolhas.

- **Comprimento Máximo da Sequência:** Não definimos um limite máximo fixo para o comprimento da sequência, utilizando o mecanismo RoPE, que permite ao modelo lidar com qualquer tamanho de janela de contexto. Nos experimentos realizados, a maior sequência processada contem 4.059 tokens, e o modelo demonstrou capacidade de extrapolar de forma eficaz para contextos desta magnitude. Essa abordagem é particularmente relevante ao trabalhar com diálogos de múltiplas etapas, permitindo que toda a sessão seja analisada sem truncamentos ou perdas de informações contextuais importantes.

- **Rank do LoRA (r)** Foi selecionado um rank LoRA de 16, indicando a dimensionalidade das matrizes de atualização de baixa ordem. Esse valor equilibra a redução do número de parâmetros e a manutenção de um rigor expressivo suficiente no modelo para capturar nuances conversacionais complexas. Um rank mais baixo arriscaria um ajuste insuficiente, enquanto um rank mais alto aumentaria os custos computacionais sem ganhos significativos de desempenho.

- **Módulos-Alvo:** O LoRA foi aplicado a todos os módulos possíveis dentro do modelo, incluindo tanto as camadas de atenção (projeções de consulta, chave, valor e saída) quanto as camadas feed-forward (projeções de entrada, saída e intermediária). Ao adaptar todos esses módulos, garantimos que o modelo capture todos os padrões e dependências relevantes necessários para um desempenho ideal em conversas de entrevista motivacional.

- **Alpha e Dropout do LoRA:** O parâmetro alpha do LoRA, definido como 16, controla a escala das atualizações de baixa ordem, garantindo estabilidade durante o treinamento ao moderar a magnitude das atualizações. Utilizamos uma taxa de dropout de 0.05, considerando que essa configuração tem se mostrado eficaz em práticas de ajuste fino de modelos grandes, como destacado em (48). Em cenários com dados de alta qualidade, taxas reduzidas de dropout podem equilibrar a regularização sem introduzir ruído desnecessário. Estudos indicam que, ao ajustar modelos utilizando LoRA, a aplicação de dropout é frequentemente menos crítica, e valores baixos como 0.05 podem evitar regularização excessiva enquanto mantêm a estabilidade do treinamento (19).

- **Bias e *Checkpointing* de Gradiente:** Não foram introduzidos vieses adicionais no modelo para evitar influenciar o processo de treinamento. O *checkpointing* de gradiente foi utilizado para otimizar o uso de memória, armazenando menos ativações intermediárias durante a retropropagação, permitindo o treinamento de modelos maiores de forma eficiente sem sacrificar o desempenho. Essa técnica é crucial para gerenciar a memória em ambientes com recursos limitados de GPU, como a instância A100 do Google Cloud utilizada em nossos experimentos.

3.5.3 Procedimento de Treinamento

O procedimento de treinamento para os modelos finais foi implementado usando o framework `transformers` da Huggingface, com hiperparâmetros selecionados de modo a garantir um treinamento eficiente e estável.

- **Tamanho do Lote e Acumulação de Gradiente:** Utilizamos um tamanho de lote de treinamento por dispositivo de 2, com passos de acumulação de gradiente de 4. Essa configuração permite acumular gradientes ao longo de múltiplos passos para aumentar efetivamente o tamanho do lote sem esgotar a memória da GPU. Essa abordagem é particularmente útil para lidar com grandes modelos, pois equilibra as restrições de memória com a necessidade de tamanhos de lote maiores para estabilizar o treinamento.

- **Passos de Aquecimento e Taxa de Aprendizado:** Foi empregada uma taxa de aprendizado de 0.0004, com 5 passos de aquecimento. Os passos de aquecimento aumentam gradualmente a taxa de aprendizado no início do treinamento para evitar instabilidade causada por grandes atualizações nas etapas iniciais do processo de otimização. Essa técnica ajuda o modelo a convergir suavemente e evita ultrapassar os parâmetros ótimos nas fases iniciais do treinamento.

- **Épocas e Passos:** O treinamento foi realizado ao longo de 230 passos, o que permitiu iterações suficientes para a convergência. A escolha de 230 passos foi determinada com base em experimentos preliminares que indicaram que o modelo atingiu bom desempenho. Esse parâmetro foi ajustado para equilibrar entre aprendizado suficiente e evitar sobrecarga computacional excessiva.

- **Otimização e Precisão:** O otimizador AdamW foi selecionado com kernels fundidos, e um decaimento de peso de 0.01 foi aplicado para prevenir o sobreajuste, penalizando pesos altos. Kernels fundidos foram empregados para otimizar o uso da GPU, reduzindo sobrecarga e acelerando o treinamento. Além disso, foi utilizado um agendador de taxa de aprendizado linear para decair a taxa de aprendizado ao longo do tempo, garantindo que o modelo faça atualizações menores à medida que se aproxima da convergência, levando a um ajuste fino mais estável.

O processo de ajuste fino envolveu o uso de conversas ou transcrições completas, incluindo anotações dos falantes, para que todo o contexto fosse utilizado durante o treinamento. O objetivo é que o modelo seja capaz de incorporar o tópico contextual associado à conversa, permitindo que suas camadas de atenção gerem respostas mais relevantes. Conforme afirmado em (102), em conversas longas, a coocorrência palavra-documento carece da informação necessária para descrever os tópicos com precisão, resultando em tópicos que se tornam muito gerais ou incoerentes.

Como demonstrado em (122), os LLMs funcionam implicitamente como modelos de variáveis latentes, onde essas variáveis capturam informações sobre a tarefa sendo executada. No caso do ajuste fino, as conversas completas foram agregadas, garantindo que todas as interações dentro de uma conversa contribuíssem para a capacidade do modelo de identificar o tópico de referência. Esse processo permite que o modelo compreenda melhor o fluxo contextual do diálogo, incluindo como os diálogos com diferentes tópicos se desenrolam, suas respectivas abordagens e particularidades. A agregação das conversas foi possível graças ao método RoPE, como mencionado anteriormente.

Entendemos que, durante o ajuste fino, o modelo aprende a associar diferentes conversas com seus respectivos tópicos, incorporando esse conhecimento em suas camadas. No momento da inferência, quando o modelo é apresentado a novas conversas, esperamos que ele seja capaz de incorporar, até certo ponto, o tópico envolvido, utilizando as relações aprendidas entre as trocas de falantes e tópicos e sua natureza semântica. Essa abordagem de ajuste fino, permite que as camadas de atenção foquem nas seções relevantes da conversa, identificando o tópico com mais precisão com base no contexto completo da conversa. Como resultado, esperamos que o modelo gere respostas que sejam não apenas contextualmente apropriadas, mas também alinhadas com o tópico específico da conversa.

Para os modelos finais, foram utilizadas 115 conversas inteiras. Selecionadas de forma aleatória a partir do subconjunto de 120 transcrições categorizadas como sendo de alta qualidade. As mensagens utilizadas para os testes foram retiradas das 5 conversas sobressalentes de alta qualidade, que não foram utilizadas no momento do treinamento.

Na próxima subseção, avaliaremos a capacidade do modelo de incorporar esses tópicos, identificando se os últimos estados ocultos de saída do modelo ajustado discriminam, e se sim, em que nível os tópicos em questão são diferenciados.

3.5.4 Avaliação da Capacidade do Modelo de Incorporar Tópicos

Para esta seção, treinamos um modelo para avaliar o quão bem os LLMs incorporam esses tópicos. Os dados possuem distribuição de tópicos dada pela Tabela 25

Apesar de termos uma distribuição heterogênea de contagem de conversas por tópico, optamos por não realizar o balanceamento ou padronização do conjunto de dados. Isso ocorre porque tais abordagens poderiam introduzir desafios adicionais e comprometer

Tabela 25 – Distribuição de tópicos dos dados

Tópico	**Diálogos**
Redução do consumo de álcool	28 (21.1%)
Cessação do tabagismo	21 (15.8%)
Perda de peso	9 (6.8%)
Tomar medicação/seguir procedimento médico	9 (6.8%)
Mais exercícios/aumento de atividade	9 (6.8%)
Redução do uso de drogas	8 (6.0%)
Redução da reincidência	7 (5.3%)
Conformidade com regras	5 (3.8%)
Gerenciamento de asma	5 (3.8%)
Gerenciamento de diabetes	5 (3.8%)
Outro	33 (24.8%)

Fonte: Elaborado pelo autor (2024).

a diversidade e a qualidade das conversas disponíveis. Por exemplo, técnicas como *under-sampling* resultariam em um número relativamente pequeno de conversas para treinamento, reduzindo significativamente a representatividade dos dados. Por outro lado, estratégias como *oversampling* ou aumento de dados (*data augmentation*), embora possam aumentar o equilíbrio, poderiam introduzir redundâncias ou viés, especialmente em tópicos sub-representados, o que potencialmente afetaria o desempenho do modelo em casos reais.

Dessa forma, preferimos preservar a distribuição original dos dados, considerando que o objetivo do modelo é aprender a lidar com uma distribuição heterogênea de tópicos, como ocorre em cenários reais. Para a análise específica nesta seção, realizamos um subconjunto do conjunto de dados, excluindo a categoria “Outro” e removendo uma conversa de cada categoria restante. Esse subconjunto contém 91 conversas e foi utilizado para avaliar nossa hipótese de forma controlada.

Fizemos o seguinte subconjunto dos dados para avaliar nossa hipótese. Para esta avaliação, excluímos a categoria “Outro“. Para cada categoria restante, removemos uma conversa, para servir de material para testes subsequentes. Isso resultou em um conjunto de dados contendo 91 conversas.

Em seguida, analisamos visualizações UMAP bidimensionais dos últimos estados ocultos de saída para os seguintes modelos:

- Gemma 7B
- Gemma 7B-finetuned
- Llama 3 8B
- Llama 3 8B-finetuned

Em que o sufixo "-finetuned" significa que o modelo foi treinado neste trabalho. A vizualização gráfica foi realizada nas mesmas 91 conversas, e como mostrado na Figura 3 dos *embeddings* gerados pelos modelos base versus os modelos ajustados. O processo de ajuste fino claramente ajuda a distinguir melhor os tópicos, com a formação de grupos bem definidos. No entanto, alguns tópicos se sobrepõem e se misturam, mostrando que, embora as camadas de atenção incorporem melhor as informações dos tópicos de forma mais discriminativa, a separação não é perfeita. Ainda assim, é evidente que o modelo começa a organizar essas informações de forma mais clara, mesmo para tópicos com menos conversas, demonstrando a capacidade do modelo de incorporar relações semânticas latentes.

O UMAP é uma técnica de redução de dimensionalidade baseada em teoria de grafos e geometria diferencial, projetada para preservar estruturas locais e globais dos dados. Ele constrói um grafo de vizinhança no espaço original e o otimiza em baixa dimensão, facilitando a visualização de padrões latentes, como os *embeddings* de modelos de linguagem (117).

De cada uma das conversas não utilizadas no treinamento dos modelos, extraímos 3 entradas de usuário, resultando num N amostral de 30. Usando os mesmos quatro modelos, incorporamos essas entradas e avaliamos a capacidade do modelo de discriminar entre tópicos com base na similaridade de cosseno e na distância euclidiana. Avaliamos a similaridade média e a diferença dentro de cada tópico, bem como a similaridade e diferença médias entre tópicos. Os resultados são mostrados abaixo na Tabela 26.

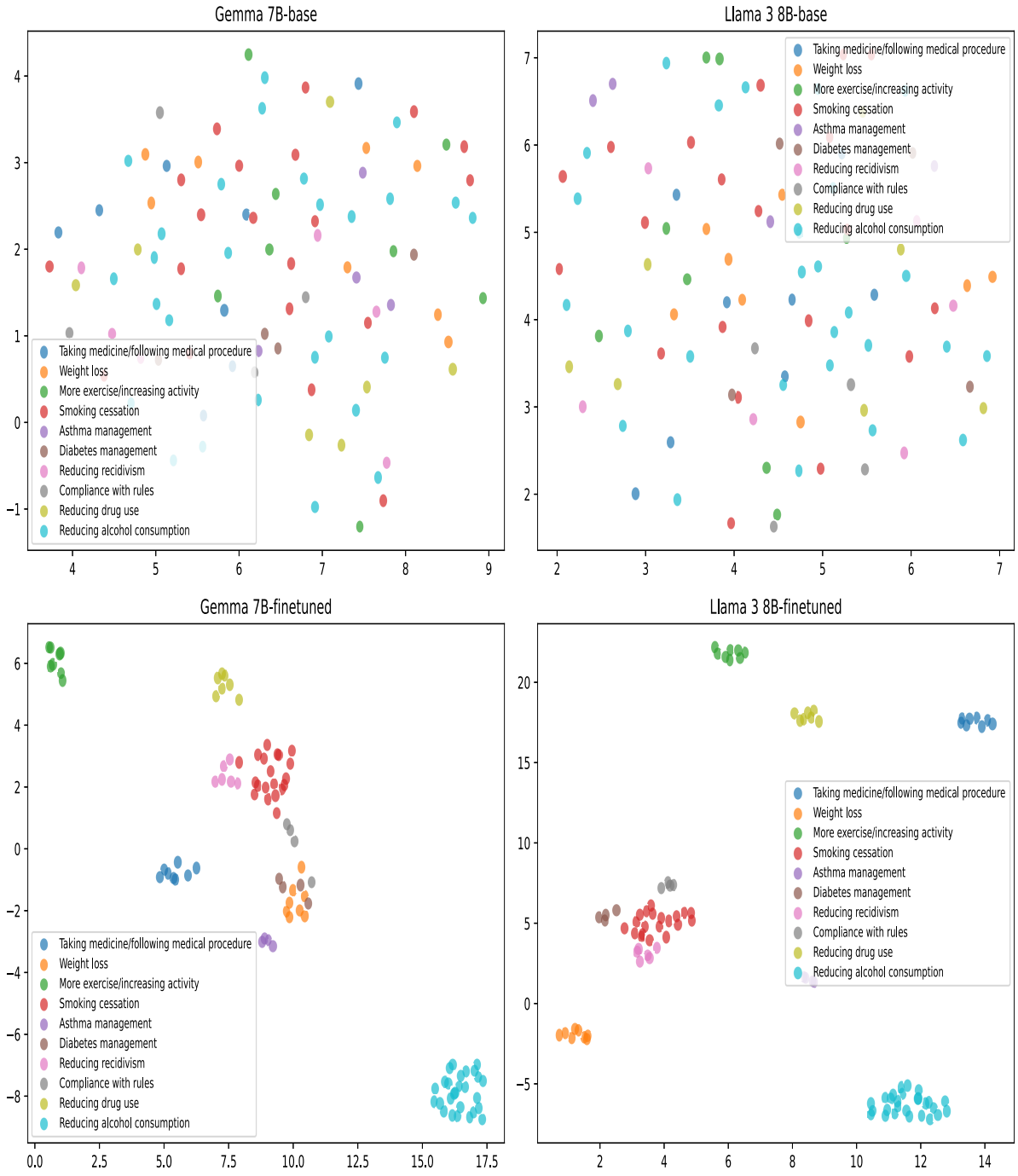
Tabela 26 – Similaridade de cosseno e distância euclidiana para comparações dentro do tópico e entre tópicos para os diferentes modelos.

Modelo	**Cosseno Dentro do Tópico**	**Cosseno Entre Tópicos**	**Euclidiana Dentro do Tópico**	**Euclidiana Entre Tópicos**
Gemma 7B-base	0.1534	0.0004	165.7252	32.5752
Llama 3 8B-base	0.1535	0.0007	115.7844	32.5899
Gemma 7B-finetuned	**0.5760**	**0.0027**	**27.5977**	**274.6720**
Llama 3 8B-finetuned	**0.5795**	**0.0021**	**27.6120**	**191.1620**

Fonte: Elaborado pelo autor (2024).

Os resultados apresentados na Tabela 26 foram submetidos a uma análise estatística para verificar se as diferenças entre os modelos base e ajustados são estatisticamente significativas. Com o tamanho da amostra de $n = 30$ para cada grupo, realizamos o teste t de Student para amostras independentes. Para a similaridade de cosseno dentro do tópico, os valores médios dos modelos ajustados ($Média = 0.5778$, $DP = 0.0025$) foram

Figura 3 – Visualizações UMAP dos últimos estados ocultos de saída.



Fonte: Elaborado pelo autor (2024).

significativamente maiores do que os dos modelos base ($Média = 0.15345$, $DP = 0.0001$), $t(58) = 92.43$, $p < 0.001$. Esses resultados confirmam que os modelos ajustados capturam significativamente melhor as nuances dentro de cada tópico. Da mesma forma, para a distância euclidiana entre tópicos, os modelos ajustados apresentaram valores médios muito maiores ($Média = 232.917$, $DP = 58.922$) em comparação com os modelos base ($Média = 32.582$, $DP = 0.0105$), $t(58) = 19.38$, $p < 0.001$. Isso indica que o ajuste fino permite

uma organização mais distinta entre tópicos. Com base nesses resultados, concluímos que o ajuste fino melhora significativamente a capacidade dos modelos de discriminar entre tópicos, tanto em termos de similaridade de cosseno quanto de distância euclidiana. Essas melhorias, no entanto, devem ser interpretadas no contexto do objetivo principal do modelo, que é não apenas distinguir tópicos, mas também manter características específicas de uma entrevista motivacional. O equilíbrio entre distinção temática e aplicação das técnicas terapêuticas é crucial para garantir interações contextualmente apropriadas e alinhadas com os princípios fundamentais da abordagem.

3.5.5 Métricas de Avaliação dos Modelos Finais

A configuração experimental consistiu nos modelos ajustados Llama 3 e Gemma, conforme detalhado na Seção **3.5.3**. Esses modelos foram então avaliados em comparação ao GPT-4, que foi utilizado como referência para a medição de desempenho. Nesse processo de avaliação, foram geradas 20 primeiras falas de usuários para representar uma ampla gama de problemas, queixas e perguntas. Essas falas foram deliberadamente elaboradas para cobrir cenários diversos, visando testar a capacidade dos modelos de lidar com diferentes tipos de consultas.

As falas dos usuários foram fornecidas a todos os modelos — Llama 3, Gemma e GPT-4 — sem qualquer contexto ou informação adicional. Essa abordagem foi adotada para avaliar o quão bem cada modelo poderia entender e responder independentemente a essas entradas isoladas. Além disso, para garantir consistência em todas as avaliações, o comprimento máximo da saída para todos os modelos, incluindo o GPT-4, foi limitado a 100 tokens. Essa restrição foi implementada para padronizar o comprimento das respostas e permitir uma comparação justa entre os modelos em termos de concisão e relevância das respostas.

O processo de avaliação incorporou tanto métricas manuais quanto automáticas. As métricas manuais foram baseadas em dois critérios distintos, anotados por três avaliadores independentes. Esses avaliadores analisaram a saída de cada modelo de diferentes perspectivas, assegurando uma avaliação mais abrangente e imparcial. Suas anotações manuais focaram em aspectos qualitativos, como a relevância e a empatia das respostas.

Paralelamente, três métricas automáticas foram empregadas para fornecer uma medida objetiva do desempenho dos modelos. Essas métricas foram definidas de acordo com as diretrizes descritas em um artigo de referência, que ofereceu uma metodologia bem estabelecida para avaliar sistemas de diálogo. As métricas automáticas ofereceram uma avaliação quantitativa, focando em aspectos como coerência, adequação emocional e fluência linguística. Ao combinar métricas manuais e automáticas, esta avaliação garantiu uma análise equilibrada e detalhada das capacidades dos modelos em gerar respostas significativas e contextualmente apropriadas.

3.5.5.1 Métricas Automáticas

As métricas automáticas de avaliação foram cuidadosamente projetadas de acordo com os princípios descritos em (105), que forneceram não apenas fundamentos teóricos, mas também implementações práticas de código. Essas métricas fazem parte do framework EPITOME, que é projetado para avaliar a comunicação empática em sistemas de diálogo.

A empatia, neste contexto, é quantificada através de vários mecanismos-chave de comunicação: Reações Emocionais, Interpretações e Explorações. Cada um desses mecanismos corresponde a um aspecto distinto da comunicação empática. As Reações Emocionais capturam a capacidade do modelo de reconhecer e responder ao tom emocional da mensagem do usuário, enquanto as Interpretações focam na habilidade do modelo de compreender e interpretar o significado ou intenção subjacente da mensagem. As Explorações avaliam a capacidade do modelo de se engajar mais profundamente com o tópico ou problema introduzido pelo usuário.

Essas métricas automáticas foram cruciais para avaliar o desempenho do modelo, particularmente na geração de respostas que são empáticas, emocionalmente ressonantes e coerentes. Ao focar nesses aspectos centrais, o framework garante que o sistema não apenas incorpora adequadamente o contexto, mas também reage de maneira apropriada a sinais emocionais, promovendo uma interação mais humanizada.

Para codificar as entradas de forma eficaz, utilizamos dois codificadores independentes de transformadores pré-treinados baseados no Roberta-base. O primeiro codificador, referido como S-*Encoder*, é utilizado para processar a mensagem do usuário, capturando suas nuances emocionais e contextuais. O segundo codificador, conhecido como R-*Encoder*, é responsável por codificar a resposta, garantindo que a resposta do modelo esteja alinhada tanto emocional quanto contextualmente com a entrada do usuário.

Além disso, os três mecanismos de comunicação — Reações Emocionais, Interpretações e Explorações — foram modelados de forma independente usando três modelos distintos dentro do framework EPITOME. Essa abordagem modular permite uma avaliação mais especializada e direcionada de cada dimensão empática, assegurando que as respostas do modelo sejam avaliadas de forma abrangente.

Para adaptar esse framework ao nosso caso específico, os conjuntos de dados usados para treinar os modelos foram traduzidos para o português. Esse passo foi crucial para manter a consistência entre os dados de treinamento e avaliação, garantindo que os cálculos das métricas refletissem com precisão a linguagem utilizada em ambos. Ao alinhar o idioma dos dados, garantimos que as capacidades empáticas do modelo sejam avaliadas em um contexto que espelha o uso no mundo real, aumentando ainda mais a confiabilidade dos resultados.

Reações Emocionais avaliam a capacidade do modelo de reconhecer e responder

ao conteúdo emocional das declarações do entrevistado. Essa métrica é calculada usando um modelo pré-treinado que identifica reações emocionais no texto, avaliando a frequência e a qualidade das respostas que reconhecem ou abordam adequadamente as emoções expressas. **Interpretações** envolvem a capacidade do modelo de fornecer interpretações perspicazes e contextualmente relevantes das declarações do entrevistado. Essa métrica mede a frequência e a qualidade das respostas interpretativas, garantindo que o modelo possa oferecer reflexões significativas que enriquecem a profundidade da conversa. Por fim, **Explorações** avaliam a eficácia do modelo em encorajar discussões e elaborações adicionais por parte do entrevistado. Essa métrica é calculada analisando a frequência e a qualidade das respostas que incentivam o entrevistado a explorar mais profundamente seus pensamentos e sentimentos, promovendo um diálogo mais rico.

3.5.5.2 Métricas Manuais

As métricas manuais propostas neste trabalho foram projetadas para capturar aspectos essenciais da entrevista motivacional e avaliar a proficiência do modelo em aplicar essas técnicas. Uma das métricas centrais é a Precisão de Escuta Reflexiva, que é fundamental para garantir que o entrevistador não apenas compreenda, mas também reflita com precisão as declarações do entrevistado. Na entrevista motivacional, a escuta reflexiva serve como uma habilidade central para promover empatia e compreensão. Cada declaração reflexiva produzida pelo modelo é avaliada quanto à sua correção e relevância.

Essa avaliação é crucial, pois reflexões que se desviam do significado pretendido pelo entrevistado podem desviar o processo terapêutico, reduzindo potencialmente a eficácia da intervenção. Para quantificar a precisão reflexiva, o número de declarações refletidas corretamente é calculado e dividido pelo número total de declarações reflexivas feitas. Essa métrica fornece uma percepção sobre o quão bem o modelo mantém a natureza reflexiva da conversa, garantindo que ele capte os pensamentos e sentimentos subjacentes do entrevistado em vez de simplesmente repetir o conteúdo superficial (75).

Paralelamente, a Razão de Perguntas Abertas é outra métrica chave, enfatizando a importância de incentivar respostas detalhadas e reflexivas do entrevistado. Perguntas abertas são vitais na entrevista motivacional, pois permitem que o entrevistado explore mais profundamente seus pensamentos e forneça respostas mais ricas e nuançadas. A razão é calculada determinando a proporção de perguntas abertas entre o número total de perguntas feitas pelo modelo. Esse tipo de questionamento promove um diálogo colaborativo, onde o entrevistado é incentivado a refletir e elaborar, em vez de ser restringido por perguntas que podem ser respondidas com um simples “sim“ ou “não“. Para o propósito desta métrica, uma pergunta é considerada aberta se exigir uma elaboração do entrevistado, promovendo assim um fluxo de conversa mais dinâmico e exploratório.

A avaliação dessas métricas foi conduzida em um ambiente experimental contro-

Tabela 27 – Resultados obtidos pelos modelos Llama, Gemma e GPT nas métricas automáticas, a saber, ER, IP e EX.

Modelo	**Reações Emocionais (ER)**			**Interpretações (IP)**			**Explorações (EX)**		
	0	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**
Llama	17	3	0	20	0	0	7	0	13
Gemma	17	3	0	20	0	0	7	0	13
GPT	19	1	0	20	0	0	18	0	2

Fonte: Elaborado pelo autor (2024).

lado. Dois especialistas em entrevista motivacional independentes foram encarregados de classificar as respostas do modelo para 20 inputs distintos. Esses casos foram anonimizados para evitar qualquer viés quanto a qual modelo gerou cada resposta. Isso garantiu que as avaliações fossem baseadas apenas na qualidade e adequação das reflexões e perguntas, em vez de qualquer preconceito sobre os modelos. Em casos em que os dois avaliadores discordaram sobre uma avaliação específica, um terceiro pesquisador independente foi consultado para resolver a divergência. Essa triangulação de avaliadores assegura uma avaliação robusta e confiável do desempenho do modelo, contribuindo para a validade geral dos achados experimentais (10).

3.6 RESULTADOS

Esta seção apresenta uma análise abrangente dos resultados obtidos ao avaliar o desempenho dos modelos Llama 3-8B, Gemma-7B e GPT-4 Turbo (83) em todas as métricas discutidas na seção anterior. Para simplificação, esses modelos são aqui referidos como Llama, Gemma e GPT, respectivamente. O GPT-4 Turbo, uma versão otimizada do GPT-4, é conhecido por oferecer maior eficiência computacional e custos reduzidos, mantendo a capacidade de gerar respostas altamente coerentes e contextualizadas.

3.6.1 Avaliação usando Métricas Automáticas

A Tabela 27 mostra os resultados das métricas automáticas. Ao examinar os resultados, é possível observar que o desempenho geral dos modelos ajustados, Llama e Gemma, foi praticamente idêntico. Essa similaridade destaca a adaptabilidade e a flexibilidade desses modelos quando expostos ao ajuste fino com dados específicos e relevantes para o domínio. O fato de ambos os modelos terem desempenhos semelhantes, apesar de suas arquiteturas distintas, indica que, quando ajustados com o mesmo conjunto de dados, suas capacidades de aprendizado convergem para resultados similares. Isso reflete o potencial desses modelos em internalizar efetivamente padrões dentro de conjuntos de dados altamente especializados.

No entanto, uma observação mais intrigante surge ao analisar o desempenho na métrica de Interpretações (IP). Nenhum dos modelos, incluindo o GPT-4, conseguiu alcançar sucesso significativo nesta categoria, com todos os modelos pontuando uniformemente no nível mais baixo. Essa uniformidade é notável, pois sugere uma limitação compartilhada em sua capacidade de interpretar aspectos mais profundos ou sutis das entradas do usuário. Um fator potencial que pode ter contribuído para esse resultado é a restrição imposta ao comprimento máximo de saída, que foi limitado a 100 tokens para todos os modelos. Essa limitação pode ter restringido a capacidade dos modelos de explorar e interpretar totalmente as sutilezas das declarações de entrada, particularmente em casos onde uma resposta mais elaborada poderia ter permitido interpretações mais ricas.

Para analisar mais a fundo o impacto do tipo de modelo nas métricas automáticas, realizamos uma análise de regressão logística. Essa abordagem estatística foi selecionada para avaliar a relação entre o tipo de modelo utilizado e a probabilidade de alcançar pontuações mais altas nas métricas automáticas restantes, especialmente Reações Emocionais (ER) e Explorações (EX). Estudos anteriores mostram que os modelos de regressão logística ordinal têm sido amplamente aplicados para lidar com variáveis dependentes ordinais, sendo adequados para cenários similares (3). Dada a falta de variação na métrica IP, ela foi excluída desta análise de regressão, pois não contribuiu com informações significativas para diferenciação.

Tabela 28 – Coeficientes, Razão de chance e p-valores da Regressão Logística Multivariada referente às métricas de avaliação automática.

Métrica	**Coeficiente**	**Gemma**	**GPT**	**Llama**
ER	Coeficiente (1)	0.0000	-1.2098	0.0000
P>|z| [0.025 0.975]		-1.736 1.736	-3.566 1.146	-1.736 1.736
	Razão de chances (1)	1.000	0.298	1.000
EX	Coeficiente (1)	0.0000	-2.8163	0.0000
P>|z| [0.025 0.975]		-1.299 1.299	-4.542 -1.090	-1.299 1.299
	Razão de chances (1)	1.000	0.060	1.000

Fonte: Elaborado pelo autor (2024).

A análise dos modelos Gemma, GPT e Llama, baseada nos resultados da regressão logística para as métricas de Reações Emocionais (ER) e Explorações (EX), fornece uma compreensão mais detalhada de como esses modelos lidam com entradas emocionalmente carregadas e estimulam a elaboração adicional nas conversas. Ao avaliar os coeficientes,

razão de chance e p-valores, podemos tirar conclusões mais precisas sobre o desempenho relativo de cada modelo.

Para a métrica de Reações Emocionais (ER), que avalia a capacidade dos modelos de reconhecer e responder adequadamente ao conteúdo emocional, tanto Gemma quanto Llama alcançam uma razão de chance de 1.000. Isso sugere que esses dois modelos são igualmente proficientes em gerar respostas emocionais apropriadas. No entanto, o GPT apresenta uma razão de chance significativamente menor de 0.298, indicando uma menor probabilidade de produzir reações emocionais adequadas em comparação com os outros modelos. Embora essa diferença de desempenho sugira uma performance inferior do GPT, os p-valores associados revelam que essas variações não são estatisticamente significativas.

Os amplos intervalos de confiança observados para todos os modelos, como [-3.566; 1.146] para o GPT, indicam uma incerteza substancial nas estimativas, o que significa que não podemos rejeitar conclusivamente a hipótese nula de ausência de diferença no desempenho dos modelos. Em outras palavras, apesar das diferenças numéricas nas razões de chances, os dados não fornecem evidências suficientes para confirmar uma disparidade estatisticamente significativa na capacidade deles de lidar com conteúdo emocional. É importante notar que a amostra de entradas analisadas foi relativamente pequena, o que limita nossa capacidade de tirar conclusões estatisticamente significativas.

Em contraste, ao examinar a métrica de Explorações (EX), que avalia a capacidade do modelo de promover o diálogo e incentivar o usuário a elaborar mais, as diferenças entre os modelos se tornam mais evidenciadas. Tanto Gemma quanto Llama mantêm uma razão de chance de 1.000, sugerindo uma probabilidade igual de promover conversas prolongadas. O GPT, no entanto, apresenta uma razão de chance consideravelmente menor de 0.060, indicando uma propensão muito menor a incentivar uma elaboração adicional durante as interações.

Os p-valores e intervalos de confiança para essa métrica revelam uma diferença estatisticamente significativa entre o GPT e os outros dois modelos. Especificamente, o desempenho do GPT é notavelmente inferior quando se trata de estimular a exploração, como refletido pelos intervalos de confiança, como [-4.542, -1.090], que sugerem fortemente uma divergência substancial dos níveis de desempenho de Gemma e Llama. Como resultado, podemos rejeitar com confiança a hipótese nula de desempenho igual para o GPT em comparação aos outros modelos em relação a essa métrica.

Essa distinção clara entre os modelos ajustados (Gemma e Llama) e o GPT provavelmente está relacionada aos conjuntos de dados nos quais esses modelos foram treinados. A capacidade de incentivar discussões e elaborações adicionais por parte do usuário está intimamente alinhada com os princípios da entrevista motivacional. Nesse contexto, não é surpreendente que os modelos treinados especificamente para essas tarefas tenham apresentado melhor desempenho.

Tabela 29 – Resultados para a Precisão de Escuta Reflexiva.

Modelo	Precisas	Imprecisas	Sem Reflexão Ativa
Llama	12	3	5
Gemma	10	8	2
GPT	0	1	19

Fonte: Elaborado pelo autor (2024).

Em resumo, enquanto os três modelos exibem níveis semelhantes de desempenho ao lidar com conteúdo emocional, conforme refletido na métrica de Reações Emocionais (ER), onde não foram observadas diferenças estatisticamente significativas, a métrica de Explorações (EX) revela um contraste mais acentuado. O GPT foi significativamente menos eficaz do o Gemma e o Llama em promover conversas prolongadas. Esses achados indicam que, embora os modelos possam lidar com conteúdo emocional com proficiência comparável, o GPT demonstra uma lacuna notável em sua capacidade de estimular uma exploração e diálogo mais aprofundados, particularmente no contexto de mensagens curtas. Isso destaca o potencial de modelos menores e especializados para superar modelos maiores em certas tarefas específicas, quando devidamente treinados.

3.6.2 Métricas Manuais

Entre os três modelos analisados (Llama, Gemma e GPT), a **Precisão de Escuta Reflexiva** mostrou variabilidade significativa (Tabela 29). As respostas foram classificadas em três categorias: *reflexões precisas*, que captam corretamente a ideia ou emoção principal da fala do interlocutor, demonstrando compreensão e empatia; *reflexões imprecisas*, que tentam interpretar a fala do interlocutor, mas apresentam erros ou distorções que não refletem adequadamente o conteúdo ou a emoção pretendida; e *instâncias sem reflexão ativa*, ocasiões em que não houve nenhuma tentativa de reflexão ativa, ou seja, nenhuma resposta se encaixou no conceito de escuta reflexiva.

O Llama obteve a maior precisão com 12 reflexões precisas, 3 reflexões imprecisas e 5 instâncias sem reflexão ativa. O Gemma seguiu com 10 reflexões precisas, 8 reflexões imprecisas e 2 instâncias sem reflexão ativa. O GPT apresentou desempenho fraco, com nenhuma reflexão precisa, 1 reflexão imprecisa e 19 instâncias sem reflexão ativa.

Com relação às **perguntas abertas** (Tabela 30), tanto o Llama quanto o Gemma fizeram 7 perguntas abertas, com 5 e 9 perguntas fechadas, respectivamente. O GPT, no entanto, não fez nenhuma pergunta aberta, 1 pergunta fechada, e em 19 instâncias nenhuma pergunta foi feita.

As métricas manuais propostas neste trabalho foram projetadas para capturar características relevantes para entrevistas motivacionais. A **Precisão de Escuta Reflexiva** e a **Razão de Perguntas Abertas** foram analisadas para os três modelos: Llama,

Tabela 30 – Resultados para a Razão de Perguntas Abertas.

Modelo	Perguntas Abertas	Perguntas Fechadas	Sem Pergunta
Llama	7	5	8
Gemma	7	9	4
GPT	0	1	19

Fonte: Elaborado pelo autor (2024).

Tabela 31 – Resíduos Padronizados para Precisão de Escuta Reflexiva.

Modelo	Precisas	Imprecisas	Sem Reflexão Ativa
Llama	2.28	-0.84	-1.27
Gemma	1.14	1.68	-1.52
GPT	-3.43	-1.05	3.75

Fonte: Elaborado pelo autor (2024).

Tabela 32 – Resíduos Padronizados para Razão de Perguntas Abertas.

Modelo	Perguntas Abertas	Perguntas Fechadas	Sem Pergunta
Llama	1.14	-0.36	-0.80
Gemma	1.14	1.32	-0.40
GPT	-2.28	-0.96	2.33

Fonte: Elaborado pelo autor (2024).

Gemma e GPT.

A métrica de **Precisão de Escuta Reflexiva** avalia o quão precisamente o entrevistador reflete as declarações do entrevistado. Na Tabela 29, o Llama alcançou a maior precisão em escuta reflexiva com 12 reflexões precisas, 3 reflexões imprecisas e 5 instâncias sem reflexão ativa. O Gemma teve 10 reflexões precisas, 8 reflexões imprecisas e 2 instâncias sem reflexão ativa. O GPT apresentou desempenho fraco, com nenhuma reflexão precisa, 1 reflexão imprecisa e 19 instâncias sem reflexão ativa. Os resultados do teste qui-quadrado ($\chi^2 = 25.76$, $p < 0.001$) indicam diferenças significativas entre os modelos.

Os resíduos padronizados na Tabela 31 destacam que o GPT teve significativamente menos reflexões precisas (-3.43) e significativamente mais instâncias sem reflexão ativa (3.75), evidenciando sua deficiência em escuta reflexiva. O desempenho do Llama foi significativamente melhor que o do GPT, com resíduos positivos para reflexões precisas (2.28) e resíduos negativos para sem reflexão ativa (-1.27). O Gemma teve resultados mistos, apresentando resíduos positivos para reflexões precisas (1.14) e imprecisas (1.68), mas resíduos negativos para sem reflexão ativa (-1.52). Isso sugere que, enquanto o Gemma é capaz de gerar declarações reflexivas, ele apresenta dificuldades com precisão

em comparação com o Llama.

Para quantificar a significância estatística da diferença entre Llama e Gemma, realizamos uma análise post-hoc comparando ambos os modelos. A diferença entre os dois, ao considerar a correção de Bonferroni para múltiplas comparações, não foi significativa.

Para a **Razão de Perguntas Abertas**, tanto o Llama quanto o Gemma fizeram 7 perguntas abertas, com 5 e 9 perguntas fechadas, respectivamente. O GPT não fez perguntas abertas, 1 pergunta fechada e teve 19 instâncias onde nenhuma pergunta foi feita (Tabela 30). Os resultados do teste qui-quadrado ($\chi^2 = 33.72$, $p < 0.001$) também indicam diferenças significativas entre os modelos.

Os resíduos padronizados na Tabela 32 mostram que o GPT teve significativamente menos perguntas abertas (-2.28) e significativamente mais instâncias sem pergunta (2.33). O desempenho do Llama foi mais equilibrado, com resíduos positivos para perguntas abertas (1.14) e resíduos negativos para perguntas fechadas (-0.36) e sem pergunta (-0.80). O Gemma, embora tenha feito o mesmo número de perguntas abertas que o Llama, teve mais perguntas fechadas (1.32) e menos instâncias sem pergunta (-0.40), indicando uma propensão a fazer mais perguntas no geral, embora com uma proporção maior sendo fechadas em comparação com o Llama.

As diferenças entre Llama e Gemma não são tão acentuadas quanto as diferenças entre esses dois modelos e o GPT. Os resíduos do teste qui-quadrado sugerem que o desempenho do GPT é significativamente pior em ambas as métricas, com uma deficiência particular em precisão de escuta reflexiva e na geração de perguntas abertas. O desempenho do Llama, embora geralmente melhor, ainda mostra espaço para melhorias, especialmente na redução do número de perguntas fechadas e no aumento da precisão em escuta reflexiva.

3.6.3 Discussão

Os resultados distinguem claramente o desempenho dos modelos menores especializados (Llama e Gemma) do modelo maior GPT em entrevistas motivacionais. As métricas manuais revelaram que Llama e Gemma superaram significativamente o GPT em precisão de escuta reflexiva e proporção de perguntas abertas. Isso indica que modelos menores e especializados são mais adequados para tarefas que exigem compreensão e engajamento mais sutis.

Llama e Gemma alcançaram pontuações mais altas em precisão de escuta reflexiva do que o GPT. Os resultados do teste qui-quadrado sugerem que as diferenças são estatisticamente significativas, com o GPT apresentando desempenho particularmente baixo. Isso pode ser atribuído ao fato de que os modelos menores foram mais focados e treinados especificamente em dados de entrevistas motivacionais, permitindo-lhes entender melhor e refletir com precisão as declarações do entrevistado.

A capacidade de fazer perguntas abertas é crítica em entrevistas motivacionais,

pois promove uma conversa mais aprofundada, e geralmente percebida como menos punitiva. Llama e Gemma fizeram significativamente mais perguntas abertas que o GPT. O desempenho do GPT nessa métrica foi particularmente deficiente, destacando uma lacuna em sua capacidade de facilitar respostas detalhadas do entrevistado.

As métricas automáticas reforçaram ainda mais as vantagens potenciais de modelos especializados menores. Llama e Gemma apresentaram desempenho semelhante ao GPT em sua capacidade de reconhecer e responder a conteúdo emocional (ER), fornecer interpretações contextualmente relevantes (IP) e incentivar uma exploração adicional (EX). No entanto, a regressão logística multivariada indicou que o tipo de modelo não influenciou significativamente o desempenho nessas métricas.

Os resultados sugerem que modelos menores e especializados podem ser ajustados em dados específicos, aprimorando seu desempenho em tarefas de nicho, e requerem menos poder computacional e memória, tornando-os mais acessíveis e econômicos.

4 ANÁLISE DAS SESSÕES TRANSCRITAS

Este capítulo apresenta uma descrição detalhada dos dados transcritos, que serão utilizados para a modelagem da intervenção. Iniciamos com uma fundamentação da intervenção baseada em habilidades de enfrentamento cognitivo-comportamentais, conforme protocolada no manual do Project MATCH (55). A natureza breve e bem estruturada desta intervenção a torna um objeto de aplicação ideal para a modelagem com o uso de LLMs, dentro das limitações e desafios inerentes a esse processo.

A seguir, iniciamos o capítulo de maneira a contextualizar teoricamente a intervenção que será posteriormente modelada, destacando suas bases e procedimentos. A abordagem cognitivo-comportamental visa modificar padrões de pensamento e comportamento disfuncionais que perpetuam o uso de substâncias, com foco no desenvolvimento de habilidades que ajudem os pacientes a manter a abstinência e a enfrentar de maneira eficaz situações de alto risco para recaídas (55).

4.1 ESTRUTURA DA INTERVENÇÃO

A intervenção baseada em habilidades de enfrentamento cognitivo-comportamentais é estruturada em um programa de 12 sessões, com 8 sessões essenciais e 4 opcionais, que são selecionadas com base nas necessidades específicas de cada paciente. Esta estrutura permite uma abordagem individualizada e flexível, fundamental para abordar a diversidade de experiências e desafios enfrentados por pessoas com dependência de álcool. A intervenção adota uma sequência lógica e cuidadosamente planejada, onde o terapeuta avalia continuamente o progresso do paciente, ajustando as sessões opcionais de acordo com os déficits de habilidades identificados, os gatilhos mais recorrentes e as áreas de maior vulnerabilidade.

As 8 sessões essenciais cobrem elementos fundamentais do tratamento, fornecendo uma base sólida de habilidades que todos os pacientes precisam desenvolver para manter a abstinência e lidar com situações de alto risco. As sessões incluem temas como a introdução ao treinamento de habilidades de enfrentamento, onde se aborda a identificação e a compreensão de situações de risco, passando pelo gerenciamento de desejos e impulsos, que ensina técnicas para reconhecer e controlar os gatilhos internos e externos que levam ao consumo de álcool. Outras sessões essenciais abordam habilidades de resolução de problemas, fornecendo uma estrutura para abordar desafios cotidianos de forma sistemática, e habilidades de recusa de bebidas, que são essenciais para lidar com a pressão social e situações onde o consumo de álcool é incentivado.

O formato das sessões segue uma estrutura consistente, que se inicia com uma revisão dos exercícios de casa e uma discussão sobre os eventos recentes na vida do paciente. Este momento inicial é crucial, pois permite que o terapeuta e o paciente explorem como

as habilidades aprendidas foram aplicadas em situações reais, identificando sucessos, dificuldades e áreas que necessitam de mais prática. Essa revisão não apenas reforça o aprendizado anterior, mas também estabelece uma conexão direta entre a terapia e os desafios enfrentados no cotidiano do paciente, aumentando o engajamento e a relevância do tratamento.

Após a revisão, cada sessão introduz uma nova habilidade ou reforça uma habilidade previamente abordada, com uma explicação clara sobre sua importância e aplicabilidade para a manutenção da sobriedade. A introdução das novas habilidades é acompanhada por demonstrações práticas conduzidas pelo terapeuta, que modela os comportamentos desejados em contextos simulados. Este processo de modelagem é um componente vital da intervenção, pois oferece ao paciente um exemplo concreto de como aplicar as habilidades, facilitando a compreensão e a internalização dos conceitos discutidos.

Um dos pilares da estrutura das sessões é o uso extensivo de *role-playing*, ou ensaio comportamental, onde o paciente pratica as novas habilidades em um ambiente controlado, com *feedback* imediato do terapeuta. O *role-playing* permite que o paciente experimente diferentes estratégias de enfrentamento em um cenário seguro, livre das pressões e consequências do mundo real. Durante esses exercícios, o terapeuta pode ajustar o nível de dificuldade das simulações, começando com situações mais simples e progredindo para desafios mais complexos conforme o paciente demonstra maior confiança e competência.

A prática supervisionada é fundamental para o desenvolvimento da autoconfiança do paciente e para a consolidação das habilidades aprendidas. Através de múltiplas repetições e ajustes baseados no *feedback* do terapeuta, o paciente refina suas respostas e aumenta sua capacidade de enfrentar situações de alto risco sem recorrer ao álcool. Essa abordagem prática também ajuda a desmistificar as situações problemáticas, permitindo que o paciente se sinta mais preparado e menos vulnerável diante dos desafios diários.

Ao final de cada sessão, são atribuídas tarefas de casa específicas que incentivam a prática contínua das habilidades discutidas durante o encontro. Essas tarefas são cuidadosamente elaboradas para serem desafiadoras, mas realizáveis, e são adaptadas para refletir os contextos e experiências únicas do paciente. A prática fora das sessões é uma parte crucial do tratamento, pois promove a generalização das habilidades para o mundo real, ajudando o paciente a transferir o que aprendeu na terapia para situações cotidianas de maneira eficaz.

A estrutura flexível da intervenção, combinada com uma abordagem sistemática de revisão, demonstração e prática supervisionada, maximiza a eficácia do tratamento. Ao proporcionar um ambiente seguro para o desenvolvimento de novas habilidades e ao oferecer oportunidades frequentes para ajustes e feedback, a intervenção promove mudanças duradouras no comportamento do paciente. Essa metodologia garante que o aprendizado

não seja apenas teórico, mas se traduza em melhorias tangíveis na capacidade do paciente de manter a sobriedade e lidar com os desafios de sua recuperação.

4.1.1 Flexibilidade e estrutura nas sessões: balanceando foco e adaptação

Embora cada sessão do programa de habilidades de enfrentamento cognitivo-comportamentais tenha um foco específico e estruturado, o manual do Project MATCH reconhece a importância de uma certa flexibilidade dentro de cada sessão. Essa flexibilidade é essencial para garantir que o tratamento seja não apenas padronizado, mas também responsivo às necessidades individuais de cada paciente, permitindo que o terapeuta adapte as intervenções de acordo com as circunstâncias únicas e os desafios vividos pelo paciente. A estrutura rígida serve como um guia para manter a coerência do tratamento, mas o manual enfatiza que o terapeuta deve estar preparado para ajustar o conteúdo e a abordagem conforme necessário, sem comprometer os objetivos centrais de cada sessão.

Cada sessão é cuidadosamente planejada para abordar um tema central, como gerenciamento de impulsos, resolução de problemas ou habilidades de recusa de bebidas. A estrutura definida de cada sessão inclui componentes essenciais como a introdução da habilidade, prática supervisionada, *role-playing* e atribuição de tarefas de casa. Esses elementos são projetados para garantir que todos os pacientes recebam uma base consistente de habilidades que são fundamentais para a manutenção da abstinência.

Apesar da importância da estrutura, o protocolo enfatiza que o terapeuta não deve se tornar excessivamente rígido ou mecânico ao seguir o protocolo. A flexibilidade dentro das sessões é vista como um aspecto essencial para atender às necessidades emergentes do paciente. Isso significa que, embora o foco da sessão deva ser mantido, o terapeuta tem a liberdade de ajustar a maneira como os conteúdos são apresentados e praticados.

Embora a flexibilidade seja uma parte integral do processo, o manual também alerta sobre a importância de manter um equilíbrio cuidadoso. A adaptação não deve desviar o foco da sessão a ponto de comprometer os objetivos terapêuticos principais. O terapeuta deve evitar que discussões pessoais ou crises desviem o tempo da sessão para longe das habilidades centrais que estão sendo trabalhadas. O manual (55) sugere que, mesmo quando a flexibilidade é aplicada, o terapeuta deve sempre tentar conectar as adaptações e discussões de volta ao tema central da sessão, utilizando as dificuldades ou crises apresentadas pelo paciente como exemplos práticos para a aplicação das habilidades.

Por exemplo, se um paciente começa a discutir um conflito recente com um familiar durante uma sessão focada em resolução de problemas, o terapeuta pode usar essa situação como um caso de estudo prático, guiando o paciente através das etapas da resolução de problemas usando o conflito real como material de trabalho. Dessa forma, o terapeuta mantém o foco na habilidade sendo desenvolvida, mas adapta a sessão para torná-la mais relevante e imediata para o paciente.

O manual do Project MATCH destaca que a flexibilidade dentro da estrutura das sessões não apenas enriquece a experiência terapêutica, mas também fortalece a eficácia do tratamento. Ao permitir que os terapeutas ajustem a abordagem para responder às necessidades individuais dos pacientes, o tratamento se torna mais engajador e impactante. A chave está em manter um equilíbrio: utilizar a estrutura para assegurar a consistência do tratamento, enquanto se adapta suficientemente para manter o foco no paciente como um indivíduo único com desafios específicos. Esta abordagem nuançada permite que o terapeuta ofereça um cuidado que é tanto cientificamente fundamentado quanto sensível às complexidades da experiência humana.

4.2 CARACTERÍSTICAS GERAIS DOS DADOS TRANSCRITOS

Nesta sessão, realizamos uma análise dos dados transcritos das sessões terapêuticas. Inicialmente, exploramos as características gerais das transcrições, buscando identificar padrões nos dados e compreender a disposição das falas. Posteriormente, examinamos a relação entre terapeuta e paciente, focando na dinâmica temporal e no conteúdo das interações. Essa análise visa não apenas melhorar o treinamento de modelos preditivos, mas também fornecer *insights* valiosos sobre o progresso e a efetividade das sessões terapêuticas.

Adicionalmente, analisamos as temáticas categorizadas previamente, destacando como os diferentes tópicos emergem nas sessões. Embora o protocolo utilizado nas intervenções siga uma estrutura padrão, observamos que a ocorrência dos tópicos varia significativamente, mesmo em um contexto estruturado. As categorias temáticas foram classificadas e analisadas, permitindo a identificação de padrões que fundamentam uma abordagem mais robusta e realista para o treinamento de modelos e intervenções futuras.

Nesta seção, apresentamos as principais características gerais dos dados utilizados no estudo. As análises subsequentes e o treinamento dos modelos baseiam-se nas transcrições de sessões de 11 pacientes. As conversas foram transcritas e diarizadas utilizando a plataforma Requalify[1]. A diarização foi posteriormente revisada manualmente, e cada fala foi classificada de forma manual com base nas categorias temáticas definidas previamente.

Na Tabela 33, apresentamos a contagem de *"tokens"* por categoria analítica em nossa base de dados. Essa distribuição fornece uma visão quantitativa das diferentes temáticas abordadas ao longo das sessões.

Observa-se que as categorias "Habilidades de recusa" e "Motivação e Entrevista Motivacional" apresentam um volume expressivamente maior de tokens, indicando que essas temáticas são fortemente abordadas nas sessões. Por outro lado, tópicos como "Resolução de problemas" e "Consciência e manejo de emoções negativas" possuem menor volume, sugerindo possíveis áreas de menor atenção ou menor necessidade percebida nas interações.

[1] https://requalify.ai

Tabela 33 – Número de tokens por categoria analítica.

Categoria	**Número de Tokens**	**Porcentagem (%)**
Habilidades de recusa	7.220.392	52.89
Motivação e Entrevista Motivacional	1.502.045	10.99
Psicoeducação sobre comorbidades	950.983	6.96
Consciência e manejo da raiva	891.092	6.53
Vantagens e desvantagens de beber	858.938	6.29
Aumentando a rede de suporte social	923.188	6.77
Atividades agradáveis	763.022	5.59
Psicoeducação sobre o tratamento	780.948	5.72
Decisões aparentemente irrelevantes	672.318	4.92
Desafios Acadêmicos e Profissionais	467.124	3.42
Planos para prevenção de recaída	142.198	1.04
Comprometimento e Responsabilidade no Tratamento	133.349	0.98
Identificação e manejo das situações de risco	135.134	0.99
Consciência e manejo de emoções negativas	128.132	0.94
Luto	120.366	0.88
Consciência e Manejo da fissura	120.332	0.88
Psicoeducação sobre o consumo	110.873	0.81
Resolução de problemas	103.092	0.76
Envolvimento com um familiar ou Relacionamento	102.994	0.75
Assertividade e habilidades de comunicação	102.981	0.75
Manejando os pensamentos permissivos	101.077	0.74

Fonte: Elaborado pelo autor (2024).

A Tabela 34 apresenta a quantidade de tópicos distintos abordados por sessão para cada paciente. Essa análise é crucial para entender a diversidade temática ao longo das sessões e como essa diversidade pode impactar o progresso terapêutico.

A análise da tabela sugere que alguns pacientes, como o Paciente A, apresentam maior diversidade de tópicos em sessões específicas, enquanto outros, como o Paciente C, têm sessões menos diversificadas. Essa variação pode refletir a individualidade nas necessidades terapêuticas e a profundidade explorada em cada sessão.

Os dados analisados até aqui destacam a relevância de categorias específicas e a dinâmica individual das sessões. Esses resultados fornecem subsídios importantes para a

Tabela 34 – Número de tópicos distintos por sessão, por paciente.

Sessão	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
0	13	7	7	6	10	-	8	11	6	10	7
1	8	10	5	6	8	-	11	7	8	11	7
2	7	11	-	7	7	-	10	8	10	1	10
3	6	7	-	6	8	-	8	12	6	6	10
4	9	10	-	8	9	-	8	6	6	6	6
5	7	10	-	4	6	-	6	-	11	10	8
6	6	9	-	5	3	-	8	-	9	9	5
7	5	8	-	6	7	-	5	-	10	8	6
8	6	10	-	2	-	8	8	-	5	9	9
9	12	8	-	7	-	11	5	-	8	9	7
10	13	-	-	5	-	8	14	-	10	11	11
11	12	-	-	2	-	11	-	-	8	8	-

Fonte: Elaborado pelo autor (2024).

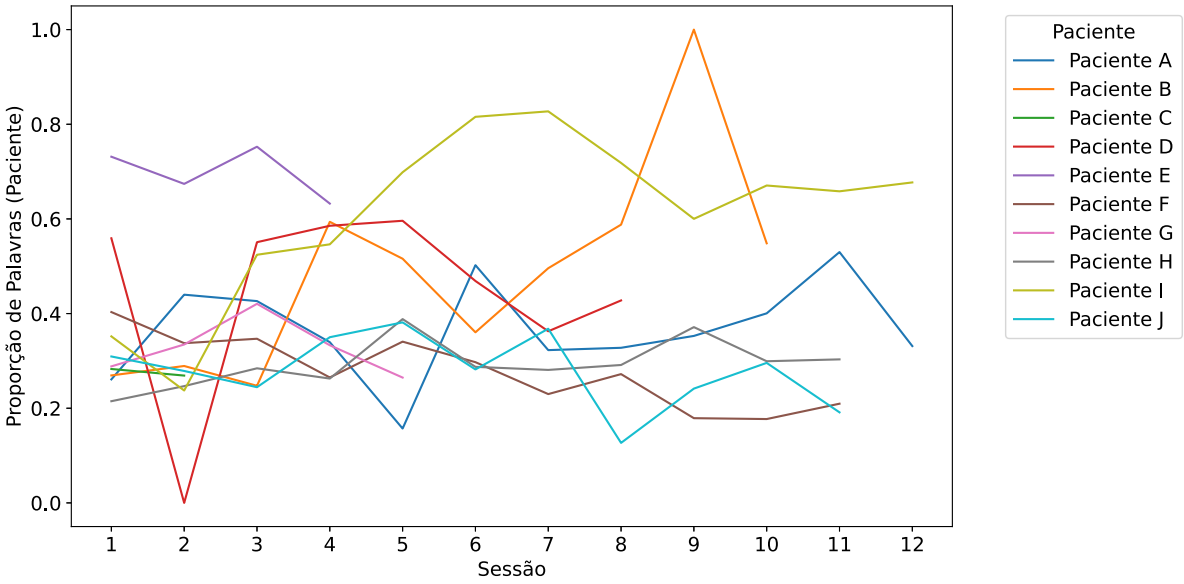
elaboração de modelos que reflitam melhor a realidade das interações terapêuticas. Nas próximas etapas, será essencial integrar análises qualitativas para compreender como os padrões numéricos observados conectam-se com os objetivos terapêuticos e os desfechos observados.

Agora, os gráficos apresentados nesta seção oferecem uma visão detalhada das dinâmicas presentes nas interações entre terapeuta e paciente ao longo das sessões. É importante destacar que, na análise gráfica, quando ocorreu interrupção da sequência de sessões de um determinado paciente, a contagem foi continuada, dado o pressuposto de que o tempo sem as seções poderia influenciar na dinâmica entre o paciente e o terapeuta. No Gráfico 4, analisamos a proporção de falas de cada falante, e observamos que existe uma tendência clara de aumento na participação do paciente conforme as sessões progridem. Esse padrão é reiterado pelo Gráfico 5, que mostra a proporção média de falas do paciente ao longo do tempo, revelando uma evolução consistente de maior engajamento. Essa dinâmica reflete o fortalecimento da aliança terapêutica, bem como o aumento da confiança e da autonomia do paciente para explorar seus próprios pensamentos, emoções e desafios durante o processo terapêutico.

O aumento na proporção de falas do paciente ao longo das sessões é particularmente relevante do ponto de vista clínico, pois sugere que as intervenções estão facilitando a criação de um espaço seguro e produtivo para o paciente. Além disso, tal evolução pode indicar que, à medida que o tratamento avança, o terapeuta está desempenhando um papel mais orientador, enquanto o paciente assume maior protagonismo nas discussões. Essa transição é essencial em muitos modelos terapêuticos, pois promove maior responsabilidade do paciente sobre seu progresso e resultados.

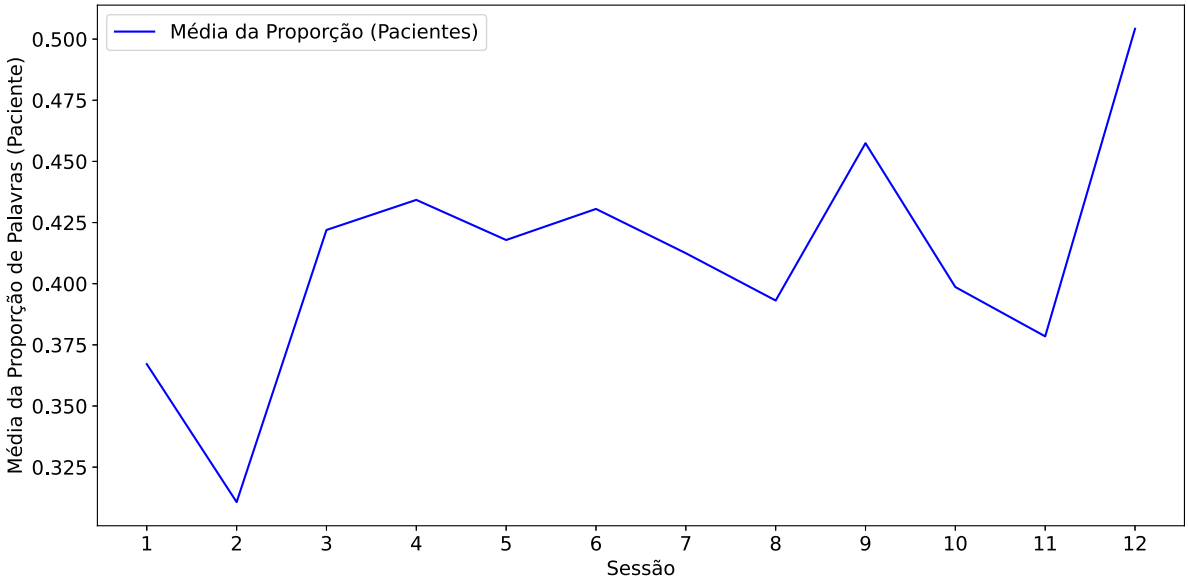
No contexto do treinamento de modelos de linguagem, essas informações oferecem

Figura 4 – Proporção de falas por falante ao longo das sessões.



Fonte: Elaborado pelo autor (2024).

Figura 5 – Proporção média de falas do paciente ao longo das seções

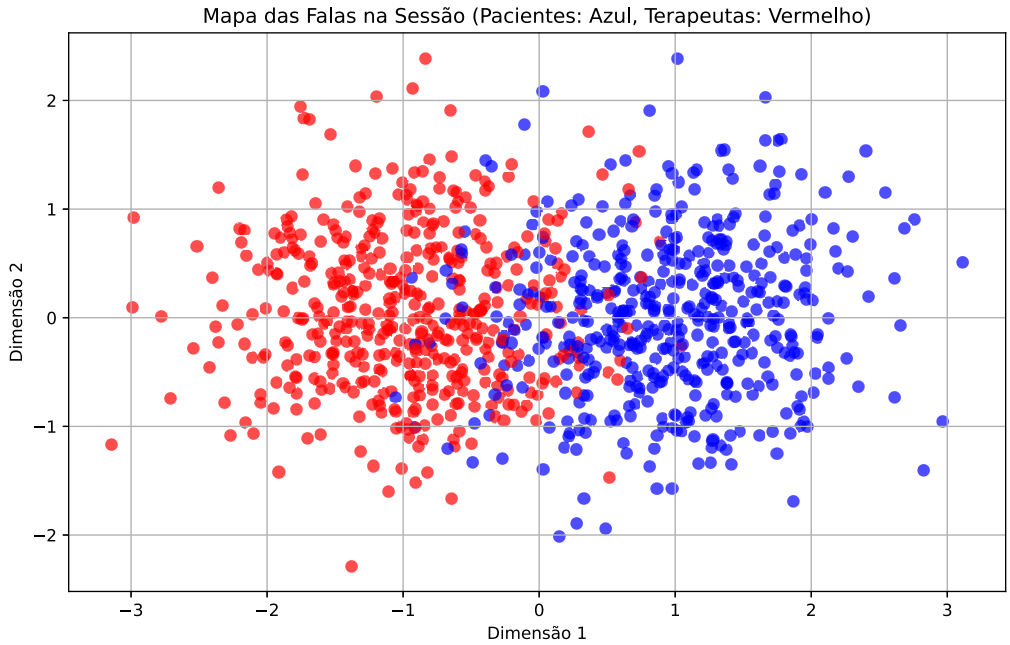


Fonte: Elaborado pelo autor (2024).

implicações valiosas. A tendência crescente na fala do paciente sugere que os modelos que visam simular ou analisar sessões terapêuticas devem ser capazes de capturar essa evolução temporal, ajustando-se à dinâmica específica de cada sessão. Isso implica que, além de modelar as interações de forma estática, o algoritmo deve ser capaz de incorporar informações sobre o estágio da interação, permitindo a adaptação a diferentes níveis de envolvimento do paciente.

A Figura 6 apresenta um mapa bidimensional das falas, destacando as transições

Figura 6 – Mapa bidimensional das falas por falante



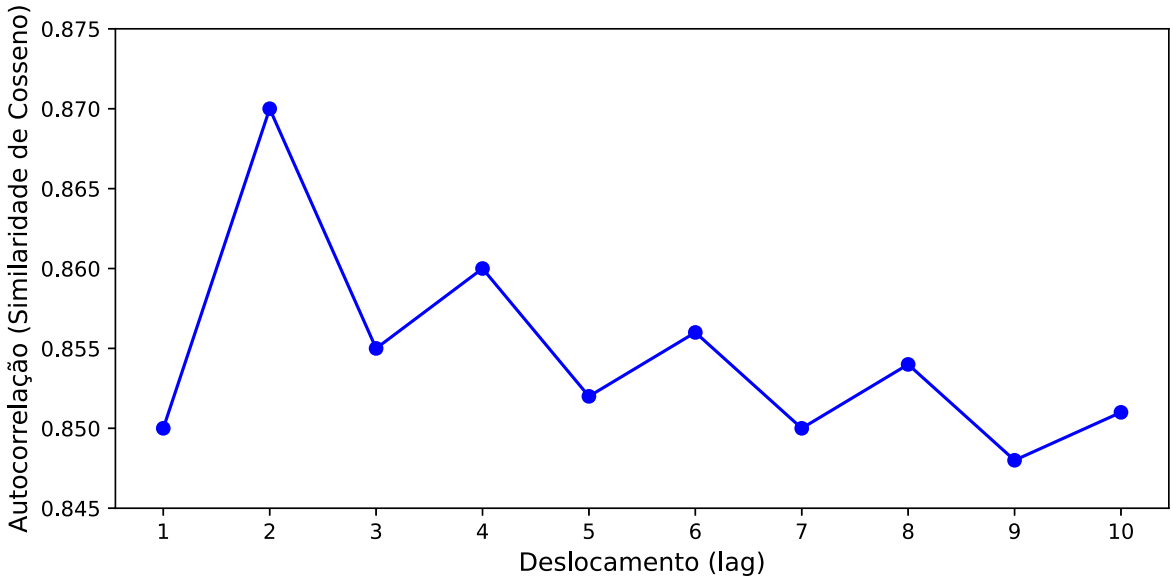
Fonte: Elaborado pelo autor (2024).

entre os dois falantes. É possível identificar uma distinção clara entre as falas do terapeuta e as do paciente, o que reforça a capacidade do modelo de reconhecer e categorizar adequadamente os participantes. Essa capacidade é fundamental para a geração de respostas contextualmente adequadas em interações simuladas, pois a identificação precisa do falante é uma das premissas básicas para manter a coerência na comunicação. A separação visual das falas também sugere que o conteúdo e o estilo de comunicação entre terapeuta e paciente possuem características distintivas, que podem ser exploradas para melhorar a personalização de modelos linguísticos, criando interações mais naturais e eficazes.

Por fim, a Figura 7 nos mostra a autocorrelação média entre as falas sequenciais, que tende a diminuir ao longo das falas, embora permaneça significativa. Essa diminuição indica que, quanto maior a distancia entre as falas, menor sua relação semântica, apresentando maior fluidez e diversificação nos tópicos abordados. Esse fenômeno é esperado, dado que muita das falas se tratam de uma continuação ou resposta direta.

Para modelos de linguagem, a redução da autocorrelação ao longo do tempo apresenta um desafio interessante. Enquanto os modelos geralmente se baseiam fortemente no contexto imediato para prever ou gerar respostas, os dados sugerem que, em interações terapêuticas, é necessário ir além do escopo local. Modelos eficazes devem ser capazes de incorporar padrões mais amplos, integrando informações de longo prazo sobre o progresso do paciente e a evolução temática da sessão. Isso não apenas melhora a precisão, mas

Figura 7 – Autocorrelção média entre as falas



Fonte: Elaborado pelo autor (2024).

também cria interações mais alinhadas à realidade da prática terapêutica.

Em síntese, a análise dos gráficos evidencia a complexidade e a riqueza das interações terapêuticas, destacando padrões significativos tanto para a compreensão do processo clínico quanto para o desenvolvimento de modelos de linguagem. A evolução na proporção de falas, a clara distinção entre os falantes e a mudança na dependência sequencial das interações reforçam a necessidade de projetar algoritmos que vão além da análise superficial da linguagem, capturando a profundidade e a dinamicidade que caracterizam as sessões terapêuticas reais.

5 PROPOSTA DE INTERVENÇÃO AUTOGUIADA

Neste capítulo, utilizaremos os dados transcritos, e o processo metodológico desenvolvido ao longo deste trabalho, para implementar uma intervenção autoguiada que replique o protocolo, de maneira automatizada. Para tanto, utilizaremos os mesmos modelos utilizados na proposta inicial. Com o uso dos mesmos parâmetros de treinamento. Em um primeiro momento, realizamos a avaliação da capacidade dos modelos de incorporar os tópicos. Depois, realizamos o treinamento, e avaliamos os modelos também de forma análoga ao que foi feito anteriormente.

5.1 AVALIAÇÃO DA CAPACIDADE DO MODELO EM INCORPORAR TÓPICOS

Nesta sessões, apresentamos o procedimento de treinamento de um (LLM) e, em seguida, avaliamos sua capacidade de incorporar tópicos. O modelo foi treinado utilizando os dados transcritos apresentados anteriormente, que incorporam os tópicos relevantes. A distribuição dos dados entre os tópicos é apresentada na Tabela 35.

Tabela 35 – Distribuição de tópicos nos dados

Tópico	**Diálogos**
Identificação e manejo das situações de risco	308 (17.3%)
Planos para prevenção de recaída	197 (11.1%)
Psicoeducação sobre o tratamento	179 (10.1%)
Consciência e manejo das emoções negativas	143 (8.0%)
Envolvimento com um familiar ou relacionamento	113 (6.4%)
Vantagens e desvantagens de beber	109 (6.1%)
Motivação e entrevista motivacional	90 (5.1%)
Atividades agradáveis	78 (4.4%)
Psicoeducação sobre o consumo	60 (3.4%)
Resolução de problemas	51 (2.9%)
Habilidades de recusa	50 (2.8%)
Assertividade e habilidades de comunicação	49 (2.8%)
Comprometimento e responsabilidade no tratamento	48 (2.7%)
Manejando os pensamentos permissivos	46 (2.6%)
Consciência e manejo da raiva	42 (2.4%)
Psicoeducação sobre comorbidades	39 (2.2%)
Decisões aparentemente irrelevantes	30 (1.7%)
Aumentando a rede de suporte social	24 (1.3%)
Desafios Acadêmicos e Profissionais	18 (1.0%)
Luto	6 (0.3%)
Total geral	1.778 (100%)

Fonte: Elaborado pelo autor (2024).

Apesar da distribuição heterogênea na quantidade de diálogos por tópico, decidiu-

se por não realizar o balanceamento ou padronização do conjunto de dados, visando preservar a distribuição natural observada nas sessões terapêuticas. Essa decisão baseia-se na premissa de que os dados refletem o contexto real das interações terapêuticas, onde tópicos mais frequentes possuem, presumidamente, maior relevância prática ou demanda de atenção nas intervenções (33). Assim, manter a proporção original dos tópicos permite que o modelo treinado esteja alinhado com a realidade observada nas sessões, garantindo maior aderência às prioridades naturais do processo terapêutico. Selecionamos um subconjunto dos dados para avaliar a capacidade dos modelos incorporarem tópicos. Para essa avaliação usamos 3 conversas de cada tópico, que não foram consideradas no treinamento do modelo, resultando em um conjunto de dados com 1718 conversas.

Analisamos visualizações bidimensionais UMAP dos estados ocultos finais para os seguintes modelos:

- Gemma 7B
- Gemma 7B-finetuned
- Llama 3 8B
- Llama 3 8B-finetuned

Para este caso, a comparação visual dos *embeddings* gerados pelos modelos base versus os modelos ajustados não é capaz de mostrar nenhuma distinção entre a capacidade de diferenciar os tópicos.

Para cada uma das 3 conversas não utilizadas no treinamento do modelo, de maneira a garantir que a avaliação seja com dados novos, usando os mesmos quatro modelos, embutimos essas entradas e avaliamos a capacidade do modelo de discriminar entre tópicos com base na similaridade do cosseno e na distância euclidiana. Avaliamos a similaridade e diferença média dentro de cada tópico, bem como a similaridade e diferença média entre tópicos. Os resultados são apresentados na Tabela 36. Neste caso, nosso N amostral é 60, dado que avaliamos 3 conversas para cada um dos tópicos.

Os resultados apresentados na Tabela 36 foram submetidos a uma análise estatística para verificar se as diferenças entre os modelos base e ajustados são estatisticamente significativas. Com o tamanho da amostra de $n = 60$ para cada grupo, realizamos o teste t de Student para amostras independentes. Para a similaridade do cosseno dentro do tópico, os valores médios dos modelos ajustados ($Média = 0.6000$, $DP = 0.0308$) foram significativamente maiores do que os dos modelos base ($Média = 0.3153$, $DP = 0.0072$), $t(118) = 56.87$, $p < 0.001$. Isso indica que os modelos ajustados capturam significativamente melhor as nuances dentro de cada tópico. Para a distância euclidiana dentro dos tópicos, os valores médios dos modelos ajustados ($Média = 3.1038$, $DP = 0.0230$) foram

Tabela 36 – Similaridade do cosseno e distância euclidiana para comparações dentro e entre tópicos para os diferentes modelos.

Modelo	Cosseno Dentro do Tópico	Cosseno Entre Tópicos	Euclidiana Dentro do Tópico	Euclidiana Entre Tópicos
Gemma 7B-base	0.3204	-0.0109	24.1453	4.3517
Llama 3 8B-base	0.3102	-0.0121	22.0996	4.3609
Gemma 7B-finetuned	0.5782	-0.0201	3.1199	45.1069
Llama 3 8B-finetuned	0.6218	-0.0119	3.0877	45.1980

Fonte: Elaborado pelo autor (2024).

significativamente menores em comparação com os dos modelos base ($Média = 23.1225$, $DP = 1.4458$), $t(118) = -81.23$, $p < 0.001$. Esses resultados refletem que o ajuste fino permite uma organização mais compacta dos *embeddings* dentro dos tópicos, criando clusters mais bem definidos. Ainda, a distância euclidiana entre tópicos mostrou um aumento significativo para os modelos ajustados ($Média = 45.1525$, $DP = 0.0645$) em comparação com os modelos base ($Média = 4.3563$, $DP = 0.0066$), $t(118) = 643.45$, $p < 0.001$. Isso reforça que o ajuste fino facilita a separação mais clara entre tópicos.

Entretanto, devido à complexidade dos tópicos, a diferenciação observada não é tão evidente, quanto no cenário anterior, demonstrado na Figura 3. Ainda assim, essa capacidade de discriminação é importante para melhorar a performance do modelo, especialmente no contexto de aplicações práticas, como garantir a consistência entre tópicos em uma abordagem de saúde mental.

5.1.1 Treinamento dos Modelos

O processo de *fine-tuning* seguido para os modelos aqui apresentados foi análogo ao protocolo utilizado na experimentação anterior. Entretanto, para este experimento, utilizaremos somente 2 modelos, ambos baseados no Llama 3-8B, dado que este mostrou maior eficiencia computacional. Iremos utilizar a mesma infraestrutura computacional e técnicas de otimização da experimentação anterior. O objetivo do treinamento foi adaptar os modelos para lidar com interações em um contexto conversacional que priorizasse precisão, fluidez e relevância das respostas.

A principal diferença entre os modelos aqui comparados reside no formato dos dados de treinamento. O modelo ajustado em pares de perguntas e respostas, referido aqui como **QA**, foi treinado com dados onde cada entrada era composta por uma única pergunta

e sua resposta correspondente. Essa abordagem segmentada enfatizou a associação entre uma entrada e sua saída correspondente. Por outro lado, o modelo ajustado em blocos de conversa, referido como **Conversa**, foi treinado com blocos inteiros de diálogo que abordavam tópicos completos. Esses blocos incluíam múltiplas perguntas, respostas e declarações contextuais, permitindo ao modelo aprender as nuances da continuidade e do encadeamento lógico de ideias ao longo de uma interação mais longa. Essa abordagem buscou capacitar o modelo a compreender e responder com maior sensibilidade ao contexto global, promovendo interações mais naturais e envolventes.

O processo de treinamento para ambos os modelos utilizou o algoritmo AdamW para otimização, com estratégias de regularização para evitar sobreajuste, como o uso de dropout. Para o ajuste fino, utilizou-se o mesmo conjunto de dados definido anteriormente. Dessa forma, assumimos que a distribuição natural dos tópicos e assuntos nas sessões é a ideal. Em estudos futuros, estratégias de curadoria e balanceamento da base devem ser exploradas.

5.1.2 Avaliação

As métricas utilizadas para a avaliação dos modelos foram as mesmas descritas na experimentação anterior. Estas métricas, abrangendo tanto análises automáticas quanto manuais, foram escolhidas para capturar a capacidade dos modelos de incorporar o contexto e responder adequadamente a interações em contextos conversacionais. Além disso, para evitar vieses na avaliação, os dados de teste foram selecionados de forma independente dos conjuntos de treinamento. Dois validadores humanos foram responsável por julgar a qualidade das respostas geradas, avaliando métricas como a **Precisão de Escuta Reflexiva** e a **Razão de Perguntas Abertas**, enquanto as métricas automáticas, como **Reação Emocional (ER)**, **Interpretação (IP)** e **Exploração (EX)**, também foram calculadas .

Essa abordagem de avaliação equilibrada foi fundamental para identificar as vantagens e limitações de cada modelo no contexto conversacional, destacando as nuances entre o foco em pares de perguntas e respostas versus blocos completos de diálogo.

5.2 RESULTADOS

Esta seção apresenta uma análise dos resultados obtidos ao avaliar o desempenho dos modelos Llama 3-8B ajustado em pares de perguntas e respostas (QA) e Llama 3-8B ajustado em blocos inteiros de conversa sobre o mesmo tópico (Conversa). Para simplificação, os modelos são aqui referidos como **QA** e **Conversa**, respectivamente.

5.2.1 Avaliação usando Métricas Automáticas

A Tabela 37 mostra os resultados das métricas automáticas.

Tabela 37 – Resultados obtidos pelos modelos QA e Conversa nas métricas automáticas, a saber, ER, IP e EX.

Modelo	**Reações Emocionais (ER)**			**Interpretações (IP)**			**Explorações (EX)**		
	0	**1**	**2**	**0**	**1**	**2**	**0**	**1**	**2**
QA	16	4	0	20	0	0	9	1	10
Conversa	14	6	0	20	0	0	6	0	14

Fonte: Elaborado pelo autor (2024).

A análise de regressão logística foi conduzida para examinar o impacto do tipo de modelo (**QA** e **Conversa**) nas métricas automáticas de **Reações Emocionais (ER)** e **Explorações (EX)**. A métrica **Interpretações (IP)** foi excluída da análise devido à ausência de variação significativa nos resultados, o que impede sua contribuição para a diferenciação estatística entre os modelos. Os coeficientes, razões de chances e intervalos de confiança associados para cada métrica estão apresentados na Tabela 38.

Os resultados indicam que, para a métrica **Reações Emocionais (ER)**, o modelo **Conversa** apresentou uma razão de chance de 1.800 em relação ao modelo **QA**. Essa razão sugere que o modelo **Conversa** é ligeiramente mais propenso a produzir respostas com níveis intermediários de reconhecimento emocional (ER=1) do que o modelo **QA**. No entanto, os valores-p associados e os amplos intervalos de confiança ([-0.478, 1.654]) indicam que essa diferença não é estatisticamente significativa. Essa falta de significância sugere que, embora o modelo **Conversa** demonstre uma tendência levemente superior nessa métrica, os desempenhos gerais dos dois modelos podem ser considerados comparáveis em termos de habilidade para reconhecer e responder ao conteúdo emocional.

Por outro lado, a análise da métrica **Explorações (EX)** revelou uma diferença mais acentuada entre os modelos. O modelo **Conversa** apresentou uma razão de chance de 3.000 em comparação ao modelo **QA**, indicando que ele é três vezes mais propenso a incentivar a elaboração adicional e promover interações mais ricas e complexas. Além disso, essa diferença foi considerada estatisticamente significativa, conforme evidenciado pelo intervalo de confiança positivo (IC: [0.235, 1.962]). Esses achados refletem a superioridade do modelo **Conversa** em explorar contextos mais amplos e construir diálogos mais aprofundados, uma capacidade possivelmente atribuída ao treinamento em blocos inteiros de conversa, que permite maior integração de informações contextuais.

A partir desses resultados, é possível inferir que, enquanto os modelos **QA** e **Conversa** apresentam desempenhos similares na métrica de **Reações Emocionais (ER)**, a métrica **Explorações (EX)** destaca um desempenho marcadamente superior do modelo **Conversa**. Essa diferença reforça a hipótese de que ajustes realizados com maior contexto durante o treinamento aprimoram a capacidade de um modelo em promover a continuidade do diálogo e incentivar a participação ativa do interlocutor. No entanto, a limitação do

tamanho da amostra e os amplos intervalos de confiança observados para algumas métricas indicam que estudos futuros com conjuntos de dados maiores e mais diversificados serão essenciais para validar e ampliar essas conclusões.

Tabela 38 – Coeficientes, razões de chances e p-valores da Regressão Logística Multivariada referente às métricas de avaliação automática.

Métrica	**Modelo de Referência**	**QA**	**Conversa**
ER	Coeficiente	0.0000	0.5878
	P>|z|	[-1.732, 2.908]	[-0.478, 1.654]
	[0.025 0.975]		
	Razão de chances	1.000	1.800
EX	Coeficiente	0.0000	1.0986
	P>|z|	[-0.978, 2.176]	[0.235, 1.962]
	[0.025 0.975]		
	Razão de chances	1.000	3.000

Fonte: Elaborado pelo autor (2024).

5.2.2 Métricas Manuais

Para avaliar as capacidades práticas dos modelos, realizamos análises estatísticas das métricas **Precisão de Escuta Reflexiva** e **Razão de Perguntas Abertas**, com o objetivo de comparar quantitativamente os desempenhos dos modelos QA e Conversa.

Os resultados para a métrica de **Precisão de Escuta Reflexiva** indicam que o modelo Conversa apresentou um número absoluto maior de reflexões precisas (13) em comparação ao modelo QA (10). Por outro lado, o modelo QA teve mais reflexões imprecisas (7 contra 5 do modelo Conversa) e mais instâncias sem reflexão ativa (3 contra 2 do modelo Conversa). Para determinar se essas diferenças são estatisticamente significativas, realizamos um teste qui-quadrado de independência, cujos resultados são apresentados abaixo.

Os valores obtidos no teste indicam que não há uma diferença estatisticamente significativa na distribuição das categorias de reflexões entre os dois modelos ($p = 0.478$). Embora o modelo Conversa demonstre uma tendência geral para maior precisão, especialmente no número de reflexões precisas, os dados analisados não são suficientes para

Tabela 39 – Teste Qui-Quadrado para Precisão de Escuta Reflexiva.

Categoria	**QA (observado)**	**Conversa (observado)**	**p-valor**
Reflexões Precisas	10	13	0.478
Reflexões Imprecisas	7	5	
Sem Reflexão Ativa	3	2	

Fonte: Elaborado pelo autor (2024).

suportar essa diferença como estatisticamente relevante. Esse resultado pode ser explicado pelo tamanho limitado da amostra, que reduz o poder estatístico do teste.

Com relação à métrica de **Razão de Perguntas Abertas**, os dados da Tabela 40 mostram que o modelo Conversa também apresentou um desempenho superior, com 10 perguntas abertas em comparação às 8 do modelo QA, além de menos instâncias sem perguntas (1 contra 3). No entanto, o modelo QA gerou mais perguntas fechadas (9 contra 7 do modelo Conversa). Para avaliar a significância dessas diferenças, utilizamos novamente o teste qui-quadrado de independência.

Tabela 40 – Teste Qui-Quadrado para Razão de Perguntas Abertas.

Categoria	**QA (observado)**	**Conversa (observado)**	**p-valor**
Perguntas Abertas	8	10	0.674
Perguntas Fechadas	9	7	
Sem Perguntas	3	1	

Fonte: Elaborado pelo autor (2024).

Os resultados para a métrica de perguntas abertas também indicaram a ausência de significância estatística entre os dois modelos ($p = 0.674$). Assim como na métrica de reflexões, o modelo Conversa apresenta um desempenho superior em termos absolutos, mas as diferenças observadas não podem ser consideradas significativas dentro do nível de confiança convencional.

Em resumo, as análises estatísticas das métricas manuais sugerem que, embora o modelo Conversa tenda a superar o modelo QA em termos de precisão em reflexões e perguntas abertas, essas diferenças não alcançam significância estatística. Esses resultados podem estar associados a limitações amostrais ou à alta variabilidade nos dados, indicando a necessidade de estudos adicionais com maior número de observações para confirmar ou refutar essas tendências. Ainda assim, os achados apontam para uma vantagem qualitativa do modelo Conversa na capacidade de gerar interações mais engajadas e relevantes.

5.2.3 Discussão

Os resultados obtidos oferecem *insighs* sobre o desempenho dos modelos ajustados para pares de perguntas e respostas (QA) e blocos inteiros de conversa (Conversa). As análises realizadas, tanto em métricas automáticas quanto manuais, destacam diferenças relevantes entre os modelos, além de algumas limitações.

Nas métricas automáticas, o modelo Conversa apresentou maior capacidade para promover interações mais ricas e complexas, evidenciada pela métrica de Explorações (EX). Este resultado está alinhado com a hipótese de que o treinamento em blocos inteiros de conversa favorece a integração de informações contextuais, possibilitando diálogos mais aprofundados.

Em relação à métrica de Reações Emocionais (ER), embora o modelo Conversa tenha mostrado uma leve vantagem, as diferenças não foram estatisticamente significativas. Isso sugere que os desempenhos gerais dos modelos são comparáveis em sua capacidade de lidar com conteúdos emocionais. Além disso, a métrica de Interpretação (IP) não apresentou variação suficiente para ser utilizada como critério diferenciador.

Nas métricas manuais, como Precisão de Escuta Reflexiva e Razão de Perguntas Abertas, o modelo Conversa obteve resultados superiores em números absolutos. No entanto, os testes estatísticos (*qui-quadrado*) não identificaram diferenças significativas, indicando que essas vantagens podem ser atribuídas ao tamanho limitado da amostra ou à alta variabilidade nos dados. Apesar disso, os resultados qualitativos sugerem que o modelo Conversa tem maior potencial para gerar interações mais engajantes e relevantes.

Em resumo, o treinamento baseado em blocos completos de conversa parece melhorar a capacidade do modelo de explorar contextos e construir diálogos mais complexos. No entanto, as limitações observadas, como o tamanho da amostra e a variabilidade nos resultados, destacam a necessidade de estudos futuros com dados mais extensos e diversificados. A inclusão de novas métricas que capturem aspectos mais sutis das interações também pode enriquecer as análises e fornecer uma avaliação mais abrangente.

6 CONCLUSÃO FINAL

No que tange a Revisão sistemática, foi possivel obter uma compreensão abrangente do cenário diversificado dos modelos de *chatbots* e suas aplicações em contextos de saúde mental, assim como da robustez das evidências que sustentam sua eficácia. Esta conclusão sintetiza os principais *insights* da análise e destaca as implicações mais amplas para o desenvolvimento e aplicação da tecnologia de *chatbots* em diferentes casos de uso, especialmente em intervenções de saúde mental.

A presente dissertação reuniu e analisou diferentes perspectivas sobre o uso de modelos de *chatbots* e tecnologias baseadas em IA, especialmente no contexto da saúde mental e em aplicações terapêuticas. A análise integrada dos três estudos apresentados revelou paralelos significativos e interdependências que oferecem *insights* sobre o papel, a eficácia e os caminhos futuros para o desenvolvimento de sistemas de IA neste campo.

Os resultados gerais destacam que o *design* e a funcionalidade dos modelos de *chatbots* devem ser adaptados às necessidades específicas dos usuários e dos contextos de aplicação. O primeiro artigo salientou que componentes como Recuperação de Conhecimento e Reconhecimento de Emoção/Sentimento apresentam eficácia distinta dependendo do alvo de saúde mental. O segundo e terceiro estudos reforçaram essa necessidade de especialização, demonstrando que modelos menores ajustados com dados específicos superam modelos gerais em tarefas que demandam empatia, precisão contextual e interação fluida.

Além disso, os resultados convergem em apontar a importância de abordagens híbridas e personalizadas para maximizar o impacto das soluções tecnológicas. Tanto no uso de componentes variados em *chatbots* para saúde mental quanto na combinação de técnicas de ajuste fino, como pares de perguntas e blocos de conversa, a integração cuidadosa dessas abordagens se mostrou crítica para alcançar melhores resultados em diferentes aplicações. A escolha entre flexibilidade e estrutura, por exemplo, emergiu como um tema transversal, destacando que a combinação de tecnologias adaptativas e estruturadas é fundamental para balancear personalização e previsibilidade em interações terapêuticas.

Outro ponto recorrente foi a relevância de se considerar fatores culturais e linguísticos no desenvolvimento de soluções tecnológicas. Modelos ajustados com dados traduzidos e abordagens específicas para idiomas e culturas sub-representadas demonstraram alto potencial, evidenciando a necessidade de superar limitações de dados em contextos globais. Isso é particularmente relevante em países de língua não inglesa, onde soluções eficazes podem expandir significativamente o acesso ao cuidado em saúde mental.

Os resultados também enfatizaram a importância da experiência do usuário, incluindo usabilidade, engajamento e *design* centrado no ser humano. Mesmo as soluções

tecnicamente sofisticadas não alcançarão seu pleno potencial se não atenderem às necessidades emocionais e psicológicas dos usuários. Essa perspectiva reforça a necessidade de um equilíbrio entre avanços técnicos e um enfoque prático, acessível e ético.

Por fim, este trabalho sugere que o futuro das aplicações de IA em saúde mental e em contextos terapêuticos reside na convergência de inovação técnica com personalização adaptativa. A combinação de componentes tecnológicos robustos, métodos de ajuste fino eficazes e uma abordagem centrada no ser humano pode levar a soluções que sejam não apenas eficientes, mas também transformadoras, proporcionando suporte acessível, empático e altamente contextualizado a populações diversas.

Com base nesses *insights*, pesquisas futuras devem explorar a integração de técnicas híbridas em maior escala, avaliar a eficácia longitudinal dessas soluções e priorizar o desenvolvimento de sistemas inclusivos que considerem as particularidades culturais e linguísticas dos usuários. Dessa forma, será possível expandir o impacto positivo da IA em áreas críticas como saúde mental e educação, promovendo interações humanas mais ricas e significativas por meio de tecnologia.

Em relação ao experimento incial, foi possivel destacar o desempenho comparativo de modelos especializados menores (Llama e Gemma) em relação a um modelo de propósito geral de grande porte, como o GPT, no contexto de entrevistas motivacionais. Os resultados indicam que, quando ajustados com dados específicos do domínio, modelos menores podem superar significativamente modelos maiores em tarefas que exigem compreensão, empatia e engajamento mais sutis.

As métricas manuais demonstram que tanto o Llama quanto o Gemma alcançaram maior precisão em escuta reflexiva e geraram uma maior proporção de perguntas abertas do que o GPT. Isso sugere que, quando o foco está nas sutilezas da comunicação que promovem interações significativas, modelos menores, quando adaptados a dados apropriados, podem estar mais bem preparados para atender a esses padrões. Os resultados do teste qui-quadrado confirmaram ainda mais que essas diferenças são estatisticamente significativas, destacando particularmente as limitações do GPT em áreas que exigem maior engajamento e personalização.

As vantagens dos modelos especializados menores incluem seu treinamento direcionado, eficiência de recursos e desempenho consistente em domínios específicos. Esses atributos os tornam particularmente valiosos em contextos clínicos e terapêuticos, onde a precisão dos padrões de comunicação é crucial.

Além disso, o fato de que modelos ajustados em dados traduzidos superaram o GPT é um achado significativo por si só. Esse resultado enfatiza a importância de desenvolver soluções para países de língua não inglesa, onde a disponibilidade limitada de dados pode ser uma restrição significativa para a construção de modelos de alto desempenho para aplicações específicas.

Pesquisas futuras devem explorar o potencial de modelos híbridos que combinem as forças de modelos menores e maiores, bem como o avanço das técnicas de ajuste fino para aprimorar ainda mais as capacidades de modelos especializados. Além disso, a incorporação de mais dados específicos do domínio pode aumentar ainda mais o desempenho geral desses modelos.

Em conclusão, este estudo enfatiza a importância de selecionar o modelo certo para a tarefa em questão. Embora modelos grandes como o GPT ofereçam capacidades amplas, modelos menores e especializados podem proporcionar desempenho superior em aplicações direcionadas, tornando-os ferramentas valiosas em campos que requerem interação detalhada e empática. Especificamente, para fins educacionais, como o treinamento de estudantes de psicologia e conselheiros não especializados, esses modelos especializados apresentam uma solução promissora para melhorar experiências de aprendizado e aprimorar habilidades práticas em entrevistas motivacionais.

Em relação a modelagem final, do agente de intervenção autoguiada,a foi realziada a comparação do desempenho de dois modelos baseados no Llama 3-8B, ajustados com abordagens distintas: pares de perguntas e respostas (**QA**) e blocos inteiros de conversa (**Conversa**). Os resultados indicam que o modelo ajustado com blocos de conversa superou consistentemente o modelo QA em tarefas que exigem maior compreensão de contexto, engajamento e fluidez em interações.

As análises demonstraram que o modelo Conversa alcançou uma **maior precisão em escuta reflexiva** e **gerou uma maior proporção de perguntas abertas** em comparação ao modelo QA. Esses resultados destacam a capacidade do modelo Conversa de captar nuances contextuais e manter uma interação mais coesa e natural. O modelo QA, embora eficaz em situações pontuais e segmentadas, mostrou limitações quando confrontado com demandas mais complexas de continuidade e sensibilidade ao contexto.

O treinamento com blocos de diálogo permitiu ao modelo Conversa aprender padrões mais ricos de interação, o que se refletiu no desempenho superior em métricas automáticas e manuais. Esses achados reforçam a importância de alinhar o ajuste fino dos modelos às características específicas das tarefas-alvo. Abordagens que integram dados com maior densidade contextual demonstraram maior eficácia em cenários que requerem interações mais humanizadas.

Em conclusão, o modelo ajustado em blocos de conversa demonstrou um desempenho superior em cenários que exigem interações complexas e contextualmente ricas, enquanto o modelo QA se destacou em situações mais diretas e segmentadas. Esses resultados enfatizam que a escolha da abordagem de treinamento deve ser guiada pelas necessidades específicas da aplicação, destacando o potencial de modelos especializados como ferramentas valiosas para interações sensíveis e detalhadas em diversos contextos.

Conforme apresentado, esta dissertação explorou o panorama diversificado dos

modelos de *chatbots* e suas aplicações, com foco na saúde mental, fornecendo uma visão abrangente das suas potencialidades, limitações e caminhos para o futuro. Os achados destacam a importância de alinhar as capacidades técnicas dos *chatbots* às necessidades específicas dos usuários e aos contextos de aplicação, reforçando que a personalização é um elemento-chave para o sucesso dessas tecnologias.

Em síntese, esta dissertação contribuiu para aprofundar o entendimento sobre os modelos de *chatbots* em saúde mental, oferecendo direções claras para avanços futuros. A combinação de inovação técnica, personalização adaptativa e inclusão cultural emerge como o caminho mais promissor para maximizar o impacto dessas tecnologias. Pesquisas futuras devem explorar a escalabilidade, a eficácia longitudinal e a integração de abordagens híbridas, garantindo que os *chatbots* sejam ferramentas não apenas tecnicamente sólidas, mas também empáticas e acessíveis. Assim, será possível consolidar o papel dos *chatbots* como aliados transformadores no cuidado à saúde mental, promovendo suporte efetivo e humanizado para uma diversidade de populações.

REFERÊNCIAS

1 Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad, Alexander Thieme, Ram Sriram, Zhongqi Yang, Yanshan Wang, Bryant Lin, et al. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *npj Digital Medicine*, 7:82, 2024.

2 A. Abd-Alrazaq et al. Effectiveness and safety of using chatbots to improve mental health. *Journal of Medical Internet Research*, 22(7):e16021, 2020.

3 Maurício Nogueira S Abreu, Adauto L Siqueira, and Waleska Teixeira Caiaffa. Regressão logística ordinal em estudos epidemiológicos. *Revista de Saúde Pública*, 43(1):183–194, 2009.

4 F. Almusharraf, J. Rose, and P. Selby. Engaging unmotivated smokers to move toward quitting: design of motivational interviewing–based chatbot through iterative interactions. *Journal of Medical Internet Research*, 22(11):e20251, 2020.

5 Fahad Almusharraf, Jonathan Rose, and Peter Selby. Engaging unmotivated smokers to move toward quitting: Design of motivational interviewing–based chatbot through iterative interactions. *Journal of Medical Internet Research*, 22(11):e20251, 2020.

6 BR Arjona, D Crisologo Jr, MI Dela Rosa, MG Montilla, RA Narvaez, and RF Suarez. Impact of digital health apps among patients with mental health issues: An integrative review. *Canadian Journal of Nursing Informatics*, 18(1), 2023.

7 Luke Balcombe. Ai chatbots in digital mental health. *Informatics*, 10(4), 2023.

8 Amit Baumel, Frederick Muench, Stav Edan, and John M Kane. Objective user engagement with mental health apps: systematic search and panel-based usage analysis. *Journal of medical Internet research*, 21(9):e14567, 2019.

9 Anna Boggiss, Nathan Consedine, Sarah Hopkins, Connor Silvester, Craig Jefferies, Paul Hofman, and Anna Serlachius. Improving the well-being of adolescents with type 1 diabetes during the covid-19 pandemic: Qualitative study exploring acceptability and clinical usability of a self-compassion chatbot. *JMIR Diabetes*, 8:e40641, 2023.

10 S. M. Boom, R. Oberink, A. J. Zonneveld, N. van Dijk, and M. R. Visser. Implementation of motivational interviewing in the general practice setting: A qualitative study. *BMC Primary Care*, 23(1), 2022.

11 Lennart Brocki, George C. Dyer, Anna Gładka, and Neo Christopher Chung. Deep learning mental health dialogue system. *arXiv preprint arXiv:2301.09412*, 2023.

12 Franziska Burger, Mark A Neerincx, and Willem-Paul Brinkman. Using a conversational agent for thought recording as a cognitive therapy task: Feasibility, content, and feedback. *Frontiers in Digital Health*, 4:930874, 2022.

13 Oscar Castro, Jacqueline Louise Mair, Alicia Salamanca-Sanabria, Aishah Alattas, Roman Keller, Shenglin Zheng, Ahmad Jabir, Xiaowen Lin, Bea Franziska Frese, Chang Siang Lim, et al. Development of "lvl up 1.0": a smartphone-based,

conversational agent-delivered holistic lifestyle intervention for the prevention of non-communicable diseases and common mental disorders. *Frontiers in Digital Health*, 5:1039171, 2023.

14 Oscar Castro, Jacqueline Louise Mair, Alicia Salamanca-Sanabria, Aishah Alattas, Roman Keller, Shenglin Zheng, Ahmad Jabir, Xiaowen Lin, Bea Franziska Frese, Chang Siang Lim, Prabhakaran Santhanam, Rob M. van Dam, Josip Car, Jimmy Lee, E Shyong Tai, Elgar Fleisch, Florian von Wangenheim, Lorainne Tudor Car, Falk Müller-Riemenschneider, and Tobias Kowatsch. Development of "lvl up 1.0": a smartphone-based, conversational agent-delivered holistic lifestyle intervention for the prevention of non-communicable diseases and common mental disorders. *Frontiers in Digital Health*, 5:1039171, 2023.

15 Benjamin Chaix, Arthur Guillemassé, Pierre Nectoux, Guillaume Delamon, Benoît Brouard, et al. Vik: A chatbot to support patients with chronic diseases. *Health*, 12(07):804, 2020.

16 Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. Pal: Persona-augmented emotional support conversation generation. *arXiv preprint arXiv:2212.09235*, 2023.

17 Emil Chiauzzi, Andre Williams, Timothy Y. Mariano, Sarah Pajarito, Athena Robinson, Andrew Kirvin-Quamme, and Valerie Forman-Hoffman. Demographic and clinical characteristics associated with anxiety and depressive symptom outcomes in users of a digital mental health intervention incorporating a relational agent. *BMC Psychiatry*, 24(79), 2024.

18 Young-Min Cho, Sunny Rai, Lyle Ungar, João Sedoc, and Sharath Chandra Guntuku. An integrative survey on mental health conversational agents to bridge computer science and medical perspectives. *Proc Conf Empir Methods Nat Lang Process*, 2023:11346–11369, 2023. Author manuscript; available in PMC 2024 Apr 12. Published in final edited form as: Proc Conf Empir Methods Nat Lang Process. 2023 Dec.

19 Hugging Face Community. Llama-13b lora alpaca. `https://huggingface.co/magicgh/llama13b-lora-alpaca`, 2023. Accessed: 2024-12-05.

20 A. Costa and F. Mendes. Calibrating long-form generations from large language models. *Journal of Artificial Intelligence Research*, 2023.

21 Hashmaryne C. Van Cuylenburg and T.N.D.S. Ginige. Emotion guru: A smart emotion tracking application with ai conversational agent for exploring and preventing depression. In *2021 International Conference on UK-China Emerging Technologies (UCET)*, pages 1–6. IEEE, 2021.

22 Kate Daley, Ines Hungerbuehler, Kate Cavanagh, Heloísa Garcia Claro, Paul Alan Swinton, and Michael Kapps. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. *Frontiers in digital health*, 2:576361, 2020.

23 Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. *arXiv preprint arXiv:2305.10172*, 2023.

24 Ismail Dergaa, Feten Fekih-Romdhane, Souheil Hallit, Alexandre Andrade Loch, Jordan M. Glenn, Mohamed Saifeddin Fessi, Mohamed Ben Aissa, Nizar Souissi, Noomen Guelmami, Sarya Swed, Abdelfatteh El Omri, Nicola Luigi Bragazzi, and Helmi Ben Saad. Chatgpt is not ready yet for use in providing mental health assessment and interventions. *Frontiers in Psychiatry*, 14:1277756, 2024.

25 Saahil Deshpande and Jim Warren. Self-harm detection for mental health chatbots. In *MIE*, pages 48–52, 2021.

26 Varshaa Dhanasekar, Yenugu Preethi, Vishali S, Praveen Joe I R, and Booma Poolan M. A chatbot to promote students mental health through emotion recognition. In *Proceedings of the Third International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 2021.

27 M. Dimenstein, Y. F. Santos, M. Brito, A. K. Severo, and C. Morais. Demanda em saúde mental em unidades de saúde da família. *Mental*, 3(5), 2005.

28 DSA. O impacto dos llms (large language models) na Área de saúde. *DSA Blog*, 2024.

29 Ahmed Fadhil, Yunlong Wang, and Harald Reiterer. Assistive conversational agent for health coaching: a validation study. *Methods of information in medicine*, 58(01):009–023, 2019.

30 Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785, 2017.

31 Tootiya Giyahchi, Sameer Singh, Ian Harris, and Cornelia Pechmann. Customized training of pretrained language models to detect post intents in online health support groups. In *Multimodal AI in Healthcare*, pages 59–73. Springer, 2022.

32 Raman Goel, Sachin Vashisht, Armaan Dhanda, and Seba Susan. An empathetic conversational agent with attentional mechanism. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4. IEEE, 2021.

33 Simon B Goldberg, Nikolaos Flemotomos, Victor R Martinez, Michael J Tanana, Patty B Kuo, Brian T Pace, Jennifer L Villatte, Panayiotis G Georgiou, Jake Van Epps, Zac E Imel, et al. Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of counseling psychology*, 67(4):438, 2020.

34 Henrique Nóbrega Grigolli, H.N.Tobin, Kleber Takashi Yoshida, Renan de Oliveira Rocha, Cibelle A. H. Amato, and Valeria Farinazzo Martins. Remi: working on the memory and logical reasoning of the elderly. In *XXII International Conference on Human Computer Interaction (Interaccion 2022)*. Association for Computing Machinery, 2022.

35 Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts. *arXiv preprint arXiv:2205.13884*, 2022.

36 Vanshika Gupta, Varun Joshi, Akshat Jain, and Inakshi Garg. Chatbot for mental health support using nlp. In *2023 4th International Conference for Emerging Technology (INCET)*, pages 1–6. IEEE, 2023.

37 Dorit Hadar-Shoval, Zohar Elyoseph, and Maya Lvovsky. The plasticity of chatgpt's mentalizing abilities: personalization for personality structures. *Frontiers in Psychiatry*, 14:1234397, 2023.

38 Sid Ahmed Hadri and Abdelkrim Bouramoul. Friendy: A deep learning based framework for assisting in young autistic children psychotherapy interventions. *Journal of Communications Software and Systems*, 19(1):30–38, 2023.

39 Riddhi Hakani, Samiksha Patil, Sakshi Patil, Siddhi Jhunjhunwala, and Khushali Deulkar. Revivify: A depression detection and control system using tweets and automated chatbot. In *2022 IEEE World Conference on Applied Intelligence and Computing (AIC)*, pages 796–801. IEEE, 2022.

40 Hee Jeong Han, Sanjana Mendu, Beth K Jaworski, Jason E Owen, and Saeed Abdullah. Ptsdialogue: designing a conversational agent to support individuals with post-traumatic stress disorder. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, pages 198–203, 2021.

41 MD Romael Haque and Sabirat Rubya. An overview of chatbot-based mobile mental health apps: Insights from app description and user reviews. *JMIR mHealth and uHealth*, 11(1):e44838, 2023.

42 Nidhin Harilal, Rushil Shah, Saumitra Sharma, and Vedanta Bhutani. Caro: An empathetic health conversational chatbot for people with major depression. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, CoDS COMAD 2020, page 349–350, New York, NY, USA, 2020. Association for Computing Machinery.

43 Abid Hassan, M. D. Iftekhar Ali, Rifat Ahammed, Sami Bourouis, and Mohammad Monirujjaman Khan. Development of nlp-integrated intelligent web system for e-mental health. *Computational and Mathematical Methods in Medicine*, 2021:1–20, 2021.

44 Sandra Hauser-Ulrich, Hansjörg Künzli, Danielle Meier-Peterhans, Tobias Kowatsch, et al. A smartphone-based health care chatbot to promote self-management of chronic pain (selma): pilot randomized controlled trial. *JMIR mHealth and uHealth*, 8(4):e15806, 2020.

45 Tanja Henkel, Annemiek J Linn, and Margot J van der Goot. Understanding the intention to use mental health chatbots among lgbtqia+ individuals: Testing and extending the utaut. In *Chatbot Research and Design: 6th International Workshop, CONVERSATIONS 2022, Amsterdam, The Netherlands, November 22–23, 2022, Revised Selected Papers*, pages 83–100. Springer, 2023.

46 Judith Hocking, Anthony Maeder, David Powers, Lua Perimal-Lewis, Beverley Dodd, and Belinda Lange. Mixed methods, single case design, feasibility trial of a motivational conversational agent for rehabilitation for adults with traumatic brain injury. *Clinical Rehabilitation*, 38(3):322–336, 2024.

47 E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

48 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Lu Wang. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2023.

49 Asma Ul Hussna, Azmiri Newaz Khan Laz, Md Shammyo Sikder, Jia Uddin, Hasan Tinmaz, and AM Esfar-E-Alam. Prerona: mental health bengali chatbot for digital counselling. In *Intelligent Human Computer Interaction: 12th International Conference, IHCI 2020, Daegu, South Korea, November 24–26, 2020, Proceedings, Part I 12*, pages 274–286. Springer, 2021.

50 Nazish Imran, Aateqa Hashmi, and Ahad Imran. Chat-gpt: Opportunities and challenges in child mental healthcare. *Pak J Med Sci*, 39(4):1191–1193, 2023.

51 Becky Inkster, Shubhankar Sarda, and Vinod Subramanian. An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106, 2018.

52 Muhammad Imran Ismael, Nik Nur Wahidah Nik Hashim, Nur Syahirah Mohd Shah, and Nur Syuhada Mohd Munir. Chatbot system for mental health in bahasa malaysia. *Journal of Integrated and Advanced Engineering (JIAE)*, 2(2):135–146, 2022.

53 G. et al. Izacard. Llama: Open and efficient foundation language models. *Disponível em: https://arxiv.org/abs/2302.13971*, 2023.

54 G. et al. Izacard. Llama: Open and efficient foundation language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

55 Ronald Kadden, Kathleen Carroll, Dennis Donovan, Ned Cooney, Peter Monti, David Abrams, Mark Litt, and Reid Hester. *Cognitive-Behavioral Coping Skills Therapy Manual: A Clinical Research Guide for Therapists Treating Individuals with Alcohol Abuse and Dependence*, volume 3 of *Project MATCH Monograph Series*. National Institute on Alcohol Abuse and Alcoholism, Rockville, MD, 2006. NIH Publication No. 94-3724, Reprinted 2006.

56 N Kavyashree and J Usha. Medibot: Healthcare assistant on mental health and well being. In *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pages 1–5. IEEE, 2023.

57 Payam Kaywan, Khandakar Ahmed, Ayman Ibaida, Yuan Miao, and Bruce Gu. Early detection of depression using a conversational ai bot: A non-clinical trial. *PLOS ONE*, 18(2):e0279743, 2023.

58 Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. Mindfuldiary: Harnessing large language model to support psychiatric patients' journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 20. ACM, 2024.

59 Florian Onur Kuhlmeier, Ulrich Gnewuch, Stefan L¨uttke, Eva-Lotta Brakemeier, and Alexander M¨adche. A personalized conversational agent to treat depression in youth and young adults–a transdisciplinary design science research project. In *International Conference on Design Science Research in Information Systems and Technology*, pages 30–41. Springer, 2022.

60 Moustafa Laymouna, Yuanchao Ma, David Lessard, Tibor Schuster, Kim Engler, and Bertrand Lebouché. Roles, users, benefits, and limitations of chatbots in health care: Rapid review. *J Med Internet Res*, 26:e56930, Jul 2024.

61 Dongkeon Lee, Kyo-Joong Oh, and Ho-Jin Choi. The chatbot feels you-a counseling service using emotional response generation. In *2017 IEEE international conference on big data and smart computing (BigComp)*, pages 437–440. IEEE, 2017.

62 Jonathan Lee, Sarah Kopelovich, Sunny Chieh Cheng, and Dong Si. Psychosis ireach: Reach for psychosis treatment using artificial intelligence. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2913–2920. IEEE, 2022.

63 Minha Lee, Sander Ackermans, Nena Van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. Caring for vincent: a chatbot for self-compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

64 Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27, 2020.

65 S.M. Lim, C.W.C. Shiau, L.J. Cheng, and Y. Lau. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: A systematic review and meta-regression. *Behavior Therapy*, 53(2):334–347, 2022.

66 X. Liu, H. Yan, C. An, X. Qiu, and D. Lin. Scaling laws of rope-based extrapolation. *arXiv preprint arXiv:2310.05209*, 2024.

67 Archana Mariappan. *Preliminary identification and therapeutic support of depression in mental health using conversational AI.* PhD thesis, Lakehead University, 2023.

68 Matthew Louis Mauriello, Nantanick Tantivasadakarn, Marco Antonio Mora-Mendoza, Emmanuel Thierry Lincoln, Grace Hon, Parsa Nowruzi, Dorien Simon, Luke Hansen, Nathaniel H Goenawan, Joshua Kim, Nikhil Gowda, Dan Jurafsky, and Pablo Enrique Paredes. A suite of mobile conversational agents for daily stress management (popbots): Mixed methods exploratory study. *JMIR Formative Research*, 5(9):e25294, 2021.

69 Thanassis Mavropoulos, Georgios Meditskos, Spyridon Symeonidis, Eleni Kamateri, Maria Rousi, Dimitris Tzimikas, Lefteris Papageorgiou, Christos Eleftheriadis, George Adamopoulos, Stefanos Vrochidis, et al. A context-aware conversational agent in the rehabilitation domain. *Future Internet*, 11(11):231, 2019.

70 Michael McTear. *Rule-Based Dialogue Systems: Architecture, Methods, and Tools*, pages 43–70. Springer, 2021.

71 Ansh Mehta, Sukhada Virkar, Jay Khatri, Rhutuja Thakur, and Ashwini Dalvi. Artificial intelligence powered chatbot for mental healthcare based on sentiment analysis. In *2022 5th International Conference on Advances in Science and Technology (ICAST)*, pages 185–189. IEEE, 2022.

72 Wondimagegn Mengist, Teshome Soromessa, and Gudina Legese. Method for conducting systematic literature review and meta-analysis for environmental science research. *MethodsX*, 7:100777, 2020.

73 T. et al. Mesnard. Gemma: Open models based on gemini research and technology. *Disponível em: https://arxiv.org/abs/2403.08295*, 2024.

74 Ridha Mezzi, Aymen Yahyaoui, Mohamed Wassim Krir, Wadii Boulila, and Anis Koubaa. Mental health intent recognition for arabic-speaking patients using the mini international neuropsychiatric interview (mini) and bert model. *Sensors*, 22(3):846, 2022.

75 W. R. Miller and S. Rollnick. *Motivational Interviewing: Helping People Change*. Guilford Press, 2012.

76 Kshitij Mishra, Priyanshu Priya, and Asif Ekbal. Help me heal: A reinforced polite and empathetic mental health and legal counseling dialogue system for crime victims. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14408–14416, 2023.

77 Mahdi Naser Moghadasi, Yu Zhuang, and Hashim Gellban. Robo: A counselor chatbot for opioid addicted patients. In *2020 2nd Symposium on Signal Processing Systems*, pages 91–95, 2020.

78 Joonas Moilanen, Niels van Berkel, Aku Visuri, Ujwal Gadiraju, Willem van der Maden, and Simo Hosio. Supporting mental health self-care discovery through a chatbot. *Frontiers in Digital Health*, 5:1034724, 2023.

79 Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of Medical Internet Research*, 20(6):e10148, 2018.

80 S. Moulya and T. R. Pragathi. Mental health assist and diagnosis conversational interface using logistic regression model for emotion and sentiment analysis. *Journal of Physics: Conference Series*, 2161, 2022.

81 Eduardo Olano-Espinosa, Jose Francisco Avila-Tomas, Cesar Minue-Lorenzo, Blanca Matilla-Pardo, María Encarnación Serrano Serrano, F Javier Martinez-Suberviola, Mario Gil-Conesa, Isabel Del Cura-González, et al. Effectiveness of a conversational chatbot (dejal@ bot) for the adult population to quit smoking: Pragmatic, multicenter, controlled, randomized clinical trial in primary care. *JMIR mHealth and uHealth*, 10(6):e34273, 2022.

82 P. Oliveira and R. Santos. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of Computational Linguistics*, 2016.

83 OpenAI. Gpt-4 turbo: Advancing efficiency and performance, 2023.

84 M Osorio, C Zepeda, and JL Carballido. Myubot: Towards an artificial intelligence agent system chat-bot for well-being and mental health. accepted to appear in proceedings artificial intelligence for health. In *PersonaLized MedIcine aNd WEllbeing Workshop at ECAI 2020*, 2020.

85 Sumit Pandey, Srishti Sharma, and Samar Wazir. Mental healthcare chatbot based on natural language processing and deep learning approaches: ted the therapist. *International Journal of Information Technology*, 14(7):3757–3766, 2022.

86 Daniel Y. Park and Hyungsook Kim. Determinants of intentions to use digital mental healthcare content among university students, faculty, and staff: Motivation, perceived usefulness, perceived ease of use, and parasocial interaction with ai chatbot. *Sustainability*, 15(872), 2023.

87 Gain Park, Jiyun Chung, and Seyoung Lee. Human vs. machine-like representation in chatbot mental health counseling: the serial mediation of psychological distance and trust on compliance intention. *Current Psychology*, 43:4352–4363, 2024.

88 SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee, and Bongwon Suh. Designing a chatbot for a brief motivational interview on stress management: Qualitative case study. *Journal of medical Internet research*, 21(4):e12231, 2019.

89 C. Pauw, R. Hill, S. Liu, and Z. Yang. Investigating the effectiveness of a conversational agent for emotional and cognitive support in reducing mental distress: A case study of emohaa. *Frontiers in Psychology*, 13:983212, 2022.

90 Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18, 2015.

91 Carmen Peuters, Laura Maenhout, Greet Cardon, Annick De Paepe, Ann DeSmet, Emelien Lauwerier, Kenji Leta, and Geert Crombez. A mobile healthy lifestyle intervention to promote mental health in adolescence: a mixed-methods evaluation. *BMC Public Health*, 24(44), 2024.

92 G. Portela. Álcool: números preocupam profissionais de saúde pública. *Fiocruz*, 2016.

93 Courtney Potts, Frida Lindström, Raymond Bond, Maurice Mulvenna, Frederick Booth, Edel Ennis, Karolina Parding, Catrine Kostenius, Thomas Broderick, Kyle Boyd, et al. A multilingual digital mental health and well-being chatbot (chatpal): Pre-post multicenter intervention study. *Journal of Medical Internet Research*, 25:e43051, 2023.

94 Vineeth R, Sukirti Maskey, Vishakan U S, and Yashpal Singh. A proposed chatbot psykh your personal therapist and stress buster using rasa open-source framework. In *2022 OPJU International Technology Conference on Emerging Technologies for Sustainable Development (OTCON)*. IEEE, 2023.

95 Komal Rani, Harshit Vishnoi, and Manas Mishra. A mental health chatbot delivering cognitive behavior therapy and remote health monitoring using nlp and ai. In *2023 International Conference on Disruptive Technologies (ICDT)*, pages 313–317. IEEE, 2023.

96 T. et al. Ribeiro. Building rag-based llm applications for production. *Journal of Computational Linguistics*, 2024.

97 Sahand Sabour, Wen Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiaxin Wen, Jialu Zhao, and Minlie Huang. A chatbot for mental health support: exploring the impact of emohaa on reducing mental distress in china. *Frontiers in Digital Health*, 5:1133987, 2023.

98 Hamid Reza Saeidnia, Marcin Kozak, Brady D Lund, and Mohammad Hassanzadeh. Evaluation of chatgpt's responses to information needs and information seeking of dementia patients. *Scientific Reports*, 14(10273), 2024.

99 T. Saha, S. Chopra, S. Saha, P. Bhattacharyya, and P. Kumar. A large-scale dataset for motivational dialogue system: An application of natural language generation to mental health. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

100 Harsh Sakhrani, Saloni Parekh, and Shubham Mahajan. Coral: An approach for conversational agents in mental health applications. *arXiv preprint arXiv:2111.08545*, 2021.

101 Intissar Salhi, Kamal El Guemmat, Mohammed Qbadou, and Khalifa Mansouri. Towards developing a pocket therapist: an intelligent adaptive psychological support chatbot against mental health disorders in a pandemic situation. *Indonesian Journal of Electrical Engineering and Computer Science*, 23(2):1200–1211, 2021.

102 S. Salmi, R. van der Mei, S. Mérelle, and S. Bhulai. Topic modeling for conversations for mental health helplines with utterance embedding. *Telematics and Informatics Reports*, 13, 2024.

103 Christine Schillings, Dominik Meißner, Benjamin Erb, Dana Schultchen, Eileen Bendig, and Olga Pollatos. A chatbot-based intervention with elme to improve stress and health-related parameters in a stressed sample: study protocol of a randomised controlled trial. *Frontiers in Digital Health*, 5:1046202, 2023.

104 Hitanshu Shah, Pravin Anilkumar, Shreya Sakpal, and Nileema Pathak. Edra–an emotional health detection and recognition assistant. In *2021 Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–6. IEEE, 2021.

105 A. Sharma, A.S. Miner, D.C. Atkins, and T. Althoff. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*, 2020.

106 Pawan Suresh Shivansh Singh, Prajjwala Yadwadkar and Vignesh. Therapy bot using machine learning for psychotherapy. *International Journal of Research in Engineering, Science and Management*, 5(6):217–218, 2022.

107 Sayed Abu Noman Siddik, B. M. Arifuzzaman, and Abul Kalam. Psyche conversa - a deep learning based chatbot framework to detect mental health state. In *2022 10th International Conference on Information and Communication Technology (ICoICT)*, pages 146–151. IEEE, 2022.

108 J. et al. Silva. An overview of chatbot-based mobile mental health apps. *Journal of Mental Health*, 2024.

109 Ben Singh, Timothy Olds, Jacinta Brinsley, Dot Dumuid, Rosa Virgara, Lisa Matricciani, Amanda Watson, Kimberley Szeto, Emily Eglitis, Aaron Miatke, et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. *npj Digital Medicine*, 6(1):118, 2023.

110 Chaitali Sinha, Saha Meheli, and Madhura Kadaba. Understanding digital mental health needs and usage with an artificial intelligence–led mental health app (wysa) during the covid-19 pandemic: Retrospective analysis. *JMIR Formative Research*, 7:e41913, 2023.

111 Diva Smriti, Tsui-Sui Annie Kao, Rahil Rathod, Ji Youn Shin, Wei Peng, Jake Williams, Munif Ishad Mujib, Meghan Colosimo, and Jina Huh-Yoo. Motivational interviewing conversational agent for parents as proxies for their children in healthy eating: development and user testing. *JMIR Human Factors*, 9(4):e38908, 2022.

112 Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *Journal of Medical Internet Research*, 21(7):e12529, 2019.

113 Michael J Tanana, Christina S Soma, Vivek Srikumar, David C Atkins, and Zac E Imel. Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *Journal of medical Internet research*, 21(7):e12529, 2019.

114 Elham Tawfik, Eman Ghallab, and Amel Moustafa. A nurse versus a chatbot – the effect of an empowerment program on chemotherapy-related side effects and the self-care behaviors of women living with breast cancer: a randomized controlled trial. *BMC Nursing*, 22:102, 2023.

115 Almira Osmanovic Thunström, Hanne Krage Carlsen, Lilas Ali, Tomas Larson, Andreas Hellström, and Steinn Steingrimsson. Usability comparison among healthy participants of an anthropomorphic digital human and a text-based chatbot as a responder to questions on mental health: Randomized controlled trial. *JMIR Human Factors*, 11:e54581, 2024.

116 Amy J. C. Trappey, Aislyn P. C. Lin, Kevin Y. K. Hsu, Charles V. Trappey, and Kevin L. K. Tu. Development of an empathy-centric counseling chatbot system capable of sentimental dialogue analysis. *Processes*, 10(930), 2022.

117 Diego Saldaña Ulloa. A process for topic modelling via word embeddings. *arXiv preprint arXiv:2312.03705*, 2023.

118 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

119 Lu Wang, Munif Ishad Mujib, Jake Williams, George Demiris, and Jina Huh-Yoo. An evaluation of generative pre-training model-based therapy chatbot for caregivers. *arXiv preprint arXiv:2107.13115*, 2021.

120 Qing Wang, Shuyuan Peng, Zhiyuan Zha, Xue Han, Chao Deng, Lun Hu, and Pengwei Hu. Enhancing the conversational agent with an emotional support system for mental health digital therapeutics. *Frontiers in Psychiatry*, 14:1148534, 2023.

121 Ruyi Wang, Jiankun Wang, Yuan Liao, and Jinyu Wang. Supervised machine learning chatbots for perinatal mental healthcare. In *2020 International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*, pages 378–383. IEEE, 2020.

122 X. Wang, W. Zhu, M. Saxon, M. Steyvers, and W. Y. Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36, 2024.

123 Z. Wu, S. Balloccu, V. Kumar, R. Helaoui, D. Reforgiato Recupero, and D. Riboni. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3):110, 2023.

124 Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. D4: a chinese dialogue dataset for depression-diagnosis-oriented chat. *arXiv preprint arXiv:2205.11764*, 2022.

125 Sakiko Yasukawa, Taku Tanaka, Kenji Yamane, Ritsuko Kano, Masatsugu Sakata, Hisashi Noma, Toshi A Furukawa, and Takuya Kishimoto. A chatbot to improve adherence to internet-based cognitive–behavioural therapy among workers with subthreshold depression: a randomised controlled trial. *BMJ Ment Health*, 27:1–7, 2024.

126 Ali Zamani, Matthew Reeson, Tyler Marshall, Mohamad Ali Gharaat, Alex Lambe Foster, Jasmine Noble, and Osmar R Zaiane. Intent and entity detection with data augmentation for a mental health virtual assistant chatbot. In *Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents*, pages 1–4, 2023.

127 Dongsong Zhang, Jaewan Lim, Lina Zhou, and Alicia A Dahl. Breaking the data value-privacy paradox in mobile mental health systems through user-centered privacy protection: A web-based survey study. *JMIR Mental Health*, 8(12):e31633, 2021.

128 Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence, 2024.

129 Asier López Zorrilla and M. Inés Torres. A multilingual neural coaching model with enhanced long-term dialogue structure. *ACM Transactions on Interactive Intelligent Systems*, 12(2):Article 16, July 2022.