



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Aplicação de LLM no Apoio Emocional: Um Estudo com o Modelo LLaMA

Guilherme Silva Souza

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador

Prof. Dr. Marcus Vinicius Lamar

Coorientador

Dr. Cláudio Mares Guia

Brasília
2025

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Coordenador: Prof. Dr. Marcelo Grandi Mandelli

Banca examinadora composta por:

Prof. Dr. Marcus Vinicius Lamar (Orientador) — CIC/UnB
Prof.^a Dr.^a Maristela Terto de Holanda — CIC/UnB
Prof. Dr. Vinicius Ruela Pereira Borges — CIC/UnB

CIP — Catalogação Internacional na Publicação

Souza, Guilherme Silva.

Aplicação de LLM no Apoio Emocional: Um Estudo com o Modelo LLaMA / Guilherme Silva Souza. Brasília : UnB, 2025.

50 p. : il. ; 29,5 cm.

Monografia (Graduação) — Universidade de Brasília, Brasília, 2025.

1. Saúde mental, 2. Chatbot, 3. Suporte emocional, 4. Modelos de linguagem natural, 5. Inteligência artificial

CDU 004

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília-DF — Brasil

Dedicatória

Dedico este trabalho à minha mãe, **Maria de Lourdes Silva Souza**, pelo amor, apoio incondicional e por sempre acreditar em mim, e ao meu pai, **Severo de Souza Neto**, que, mesmo não estando mais presente fisicamente, continua vivo em minhas lembranças e em cada conquista da minha vida. Este trabalho é, em parte, reflexo dos valores e ensinamentos que recebi de vocês. Obrigado por tudo.

Agradecimentos

A jornada para a realização deste trabalho não teria sido possível sem o apoio, a orientação e a presença de pessoas especiais, às quais expresso minha mais sincera gratidão.

Primeiramente, agradeço ao meu orientador, **Marcus Vinicius Lamar**, por sua dedicação em me guiar durante todo o processo. Sua paciência e empatia nos momentos difíceis foram fundamentais para que eu pudesse seguir em frente com confiança.

Ao meu coorientador, **Dr. Cláudio Mares Guia**, minha gratidão por sempre acreditar no projeto e por suas contribuições valiosas, trazendo um olhar profissional da área da saúde que enriqueceu imensamente este trabalho.

Também agradeço, de coração, à minha melhor amiga, **Maria Eduarda Machado de Holanda**, a pessoa mais incrível que conheci neste curso. A pandemia tornou essa jornada ainda mais desafiadora, e compartilhar matérias, enfrentar os desafios acadêmicos juntos e desenvolver cada projeto lado a lado foram pilares de força que me ajudaram a chegar até aqui. Sou imensamente grato por cada conversa, cada jogatina e cada abraço, que tornaram os dias mais leves e essa caminhada muito mais significativa.

Por fim, a todos que, de alguma forma, fizeram parte desta caminhada, meu mais profundo agradecimento.

Resumo

Nas últimas décadas, o número de pessoas com transtornos de saúde mental, como ansiedade e depressão, aumentou significativamente, agravado por fatores como a pandemia de COVID-19 e as dificuldades de acesso a serviços psicológicos. Paralelamente, os avanços na inteligência artificial têm possibilitado um grau cada vez maior de empatia nas interações com *chatbots*, tornando-os ferramentas promissoras para apoio emocional. Diante desse cenário, este trabalho apresenta o desenvolvimento do Guiga, um aplicativo digital de saúde mental que utiliza um *chatbot* baseado no modelo *LLaMA 3*, projetado para oferecer suporte emocional contínuo a qualquer pessoa que precise ser ouvida ou desabafar. A proposta destaca-se pela acessibilidade e empatia da solução, considerando a crescente demanda por apoio psicológico e as barreiras de acesso aos serviços tradicionais. O estudo inclui a adaptação do modelo de linguagem por meio de *fine-tuning*, ajustando suas respostas para melhor atender às necessidades emocionais dos usuários. A avaliação do sistema foi realizada por meio de interações com usuários reais, que testaram tanto o modelo base quanto o modelo ajustado, seguido de um questionário para medir a percepção da experiência. Como resultado, busca-se demonstrar que o modelo ajustado proporciona interações de maior qualidade e um suporte emocional mais eficaz. Os achados deste estudo contribuem para o aprimoramento de ferramentas digitais voltadas à saúde mental, incentivando pesquisas futuras sobre personalização e eficácia de *chatbots* nesse contexto.

Palavras-chave: Saúde mental, Chatbot, Suporte emocional, Modelos de linguagem natural, Inteligência artificial

Abstract

In recent decades, the number of people experiencing mental health disorders, such as anxiety and depression, has significantly increased, exacerbated by factors like the COVID-19 pandemic and limited access to psychological services. At the same time, advancements in artificial intelligence have enabled a greater degree of empathy in chatbot interactions, making them promising tools for emotional support. In response to this scenario, this study presents the development of Guiga, a digital mental health application that integrates a chatbot based on the LLaMA 3 model, designed to provide continuous emotional support to anyone in need of being heard or venting. The proposal stands out for its accessibility and empathetic approach, considering the growing demand for psychological support and the barriers to accessing traditional services. The study includes the adaptation of the language model through fine-tuning, refining its responses to better address users' emotional needs. The system's evaluation was conducted through interactions with real users, who tested both the base and fine-tuned models, followed by a questionnaire to assess their experience. The expected outcome is to demonstrate that the fine-tuned model offers higher-quality interactions and more effective emotional support. The findings of this study contribute to the improvement of digital mental health tools, encouraging further research on the personalization and effectiveness of chatbots in this domain.

Keywords: Mental health, Chatbot, Emotional support, Language models, Artificial intelligence

Sumário

1	Introdução	1
1.1	Problema	1
1.2	Hipótese	2
1.3	Objetivos	2
1.4	Organização do Trabalho	2
2	Estado da Arte	4
2.1	Abordagens Psicológicas e a Escuta Ativa	4
2.1.1	Psicanálise	4
2.1.2	Terapia Cognitivo-Comportamental (TCC)	5
2.1.3	Abordagem Humanista	5
2.1.4	A Escuta Ativa como Pilar do Guiga	5
2.2	Histórico dos <i>Chatbots</i> para Saúde Mental	5
2.3	Aplicativos Atuais para Saúde Mental	6
2.4	Artigos Relacionados	7
3	Modelos de Linguagem	8
3.1	Modelos LLaMA	9
3.2	Fine-tuning	11
4	Metodologia Proposta	14
4.1	Processo do Fine-tuning	14
4.2	Arquitetura do Sistema	16
4.3	Telas do Aplicativo	17
4.3.1	Tela de Login	17
4.3.2	Tela de Introdução	18
4.3.3	Tela de Chat	19
4.4	<i>Prompt Engineering</i>	20
4.5	Estrutura do Aplicativo	21
4.6	Firestore	22

4.7	Ollama e LLaMA.cpp	22
4.8	Zrok	23
5	Resultados Obtidos	24
5.1	Resultados do Fine-tuning	24
5.2	Testes com Usuários Reais	27
5.3	Exemplos de Interações	28
5.3.1	Comparação com o <i>Chat GPT</i> e o Eliza	29
5.3.2	Exemplo com o modelo <i>fine-tuned</i>	32
5.4	Considerações Finais	32
6	Conclusões	34
6.1	Trabalhos Futuros	35
	Referências	36
	Anexo	38
I	Formulários Aplicados aos Usuários	39

Lista de Figuras

3.1	Arquitetura do modelo LLaMA 3	10
3.2	Fluxo de treinamento de um modelo de linguagem.	11
3.3	Comparação entre diferentes técnicas de ajuste fino de modelos de linguagem.	12
4.1	Exemplo de entrada do <i>dataset</i> formatado.	15
4.2	Arquitetura do sistema, ilustrando a interação entre o aplicativo, o servidor de inferência <i>Ollama</i> e o banco de dados <i>Firebase Firestore</i>	17
4.3	Tela de <i>login</i> do aplicativo.	18
4.4	Tela de introdução.	19
4.5	Tela de chat do aplicativo.	20
5.1	Evolução do Gradiente das Normas ao longo dos passos.	25
5.2	Variação da taxa de aprendizado ao longo do treinamento.	25
5.3	Evolução da função de perda ao longo dos passos de treinamento.	26
5.4	Perda suavizada ao longo dos passos de treinamento (janela de 100 passos).	27
I.1	Formulário sobre o modelo base parte 1	40
I.2	Formulário sobre o modelo base parte 2	41
I.3	Formulário sobre o modelo com FT parte 1	42
I.4	Formulário sobre o modelo com FT parte 2	43

Lista de Tabelas

4.1	Exemplo de <i>Prompt Engineer</i> aplicado no Guiga	21
5.1	Comparativo Quantitativo entre os Modelos Base e <i>Fine-Tuned</i>	28

Lista de Abreviaturas e Siglas

API Interface de Programação de Aplicações (do inglês *Application Programming Interface*).

FT Ajuste Fino (do inglês *Fine-tuning*).

GPU Unidade de Processamento Gráfica (do inglês *Graphics Processing Unit*).

IA Inteligência Artificial.

LLaMA *Large Language Model Meta AI*.

LLM Modelo de Linguagem em Grande Escala (do inglês *Large Language Model*).

LoRA *Low-Rank Adaptation*.

NLP processamento de linguagem natural (do inglês *Natural Language Processing*).

QLoRA *Quantization-aware Low-Rank Adaptation*.

SFT Ajuste Fino Supervisionado (do inglês *Supervised Fine-Tuning*).

TCC Terapia Cognitivo-Comportamental.

UnB Universidade de Brasília.

Capítulo 1

Introdução

1.1 Problema

Nos últimos anos, houve um aumento expressivo na conscientização e na necessidade de soluções eficazes para problemas relacionados à saúde mental. A pandemia de COVID-19 intensificou essa preocupação, resultando em um aumento substancial nos casos de depressão, ansiedade e insônia em várias populações ao redor do mundo. Lakhan *et al.* [1] relatou um crescimento de 20% nos casos de depressão e 35% nos de ansiedade entre 113.285 indivíduos em 16 estudos realizados durante o período inicial da pandemia. Além disso, uma meta-análise internacional [2] realizada durante a pandemia de COVID-19 revelou que a prevalência de sintomas de ansiedade e insônia variou consideravelmente entre os estudos, estando frequentemente na faixa de 15% a 20% em populações adultas. Esses transtornos impactam negativamente a vida diária, prejudicando a funcionalidade social e aumentando os riscos de autolesões e suicídio.

Embora a demanda por apoio psicológico esteja crescendo, os sistemas de saúde mental enfrentam desafios significativos, como o estigma associado às doenças mentais e a dificuldade de acesso a serviços adequados, especialmente devido aos custos elevados das terapias e à concentração de serviços em grandes centros urbanos [3, 4]. Esse cenário contribui para diagnósticos tardios e tratamentos insuficientes, agravando a situação dos pacientes.

A adoção de tecnologias digitais, como aplicativos de saúde mental, surge como uma alternativa promissora para preencher essa lacuna, oferecendo suporte inicial e contínuo a baixo custo. No entanto, a personalização e a adequação cultural dessas ferramentas permanecem como barreiras críticas a serem superadas para garantir sua eficácia e impacto positivo [5].

1.2 Hipótese

Este estudo parte da hipótese de que um aplicativo digital com um *chatbot* baseado em Modelo de Linguagem em Grande Escala (do inglês *Large Language Model*) (LLM) pode oferecer suporte emocional eficaz e contínuo a qualquer pessoa que precise ser ouvida, esteja se sentindo sozinha ou necessite desabafar. Historicamente, a utilização de *chatbots* para apoio emocional remonta à década de 1960, com a criação do programa *Eliza*[6], uma tentativa rudimentar de simular uma interação terapêutica. Desde então, a evolução das tecnologias de inteligência artificial possibilitou o desenvolvimento de sistemas mais avançados, capazes de compreender melhor as necessidades dos usuários e fornecer um apoio mais empático e personalizado.

1.3 Objetivos

Este estudo tem como principal objetivo explorar o potencial do Guiga, um aplicativo de saúde mental que integra um *chatbot* projetado para oferecer apoio emocional a qualquer pessoa que precise ser ouvida, esteja enfrentando a solidão ou necessite expressar seus sentimentos. A personificação do *chatbot* com o nome Guiga busca humanizar a interação, proporcionando uma experiência mais próxima de uma conversa com um ser humano. A proposta do aplicativo é desenvolver uma solução acessível, garantindo facilidade de uso para uma ampla gama de usuários; empática, promovendo interações mais humanas e personalizadas por meio de respostas naturais; e baseada em evidências, fundamentada em pesquisas científicas para assegurar a eficácia no suporte emocional.

Além disso, busca-se aprimorar a qualidade das interações do sistema proposto por meio do processo de Ajuste Fino (do inglês *Fine-tuning*) (FT) do modelo base, ajustando suas respostas para torná-las mais alinhadas às necessidades emocionais dos usuários. Ao final do estudo, espera-se contribuir para a melhoria contínua de aplicativos voltados à saúde mental, destacando a importância de uma abordagem centrada no usuário. Com essa integração avançada, pretende-se oferecer uma ferramenta prática para o gerenciamento da saúde mental, promovendo o autocuidado e a qualidade de vida dos indivíduos afetados por essa condição.

1.4 Organização do Trabalho

Este trabalho foi estruturado de maneira a apresentar de forma clara e organizada os aspectos fundamentais do estudo, dividindo-o em seis capítulos principais.

Inicialmente, no Capítulo 2 Estado da Arte, é feito um levantamento sobre o uso de *chatbots* voltados para apoio emocional e saúde mental, analisando aplicações já existentes e discutindo estudos realizados na área.

O Capítulo 3, Modelos de Linguagem, apresenta uma introdução aos LLMs, com foco nos modelos da família *LLaMA*. São abordadas suas diferentes versões e avanços, além do processo de FT realizado para a adaptação ao domínio de suporte emocional.

No Capítulo 4, Metodologia Proposta, são descritos os detalhes técnicos do desenvolvimento da solução, incluindo o uso do *framework Flutter* para a criação do aplicativo. No capítulo, é mostrado o experimento realizado com o FT, detalhando sua configuração de treinamento, incluindo hiperparâmetros e o ambiente onde foi conduzido. Além disso, são apresentadas a estrutura do aplicativo, a integração com o *Firestore* para armazenamento de dados e autenticação de usuários, bem como a conexão com o modelo de linguagem via *Ollama* e *Zrok*, permitindo a comunicação entre o aplicativo e o servidor de inferência do modelo.

No Capítulo 5, Resultados Obtidos, são apresentados os resultados do FT com base no comportamento da perda ao longo do processo, nas normas dos gradientes e na taxa de aprendizado, que servem como métricas avaliadoras. Além disso, são realizadas análises quantitativas e qualitativas da comparação entre os modelos, incorporando métricas obtidas a partir das interações com usuários reais.

Por fim, o trabalho se encerra com o Capítulo 6 de Conclusões, onde são discutidas as principais contribuições e possíveis direções para pesquisas futuras. O estudo é finalizado com a seção de Referências, onde estão reunidas as fontes utilizadas ao longo do desenvolvimento do trabalho.

Capítulo 2

Estado da Arte

Neste capítulo, apresentamos os fundamentos teóricos que embasam o uso de *chatbots* no contexto da saúde mental. Iniciamos com uma contextualização das principais abordagens terapêuticas. Em seguida, traçamos um panorama histórico, abordando a evolução dessas tecnologias desde os primeiros sistemas especialistas, como o *Eliza*, até os modelos contemporâneos baseados em Inteligência Artificial (IA). Posteriormente, discutimos os aplicativos atuais voltados para a saúde mental, destacando aqueles que utilizam *chatbots* para oferecer suporte emocional mais dinâmico e interativo. Por fim, revisamos estudos acadêmicos relevantes que exploram a eficácia dessas tecnologias, fornecendo uma base teórica para compreender seu impacto na promoção da saúde mental.

2.1 Abordagens Psicológicas e a Escuta Ativa

Antes de falarmos sobre *chatbots*, precisamos contextualizar as abordagens terapêuticas. A psicoterapia é um campo amplo que abarca diversas escolas teóricas, cada uma com suas próprias metodologias e princípios fundamentais. Para o desenvolvimento do aplicativo Guiga, destacamos algumas abordagens psicológicas relevantes, ressaltando que o foco principal será na escuta ativa.

2.1.1 Psicanálise

A psicanálise, proposta por Sigmund Freud, tem como princípio a exploração do inconsciente e a interpretação dos conteúdos reprimidos [7]. O terapeuta atua como um facilitador, ajudando o paciente a acessar suas emoções e memórias reprimidas. Essa abordagem se baseia na associação livre e na interpretação de sonhos como formas de acessar conteúdos inconscientes.

2.1.2 Terapia Cognitivo-Comportamental (TCC)

A Terapia Cognitivo-Comportamental é uma abordagem estruturada e baseada em evidências que foca na modificação de padrões de pensamento e comportamento disfuncionais. Criada por *Aaron Beck* e *Albert Ellis*, a TCC utiliza técnicas como reestruturação cognitiva e exposição gradual para tratar diversas condições psicológicas, incluindo depressão e ansiedade [8].

2.1.3 Abordagem Humanista

A abordagem humanista, com destaque para a Terapia Centrada na Pessoa de Carl Rogers, enfatiza a importância da empatia, autenticidade e aceitação incondicional. O terapeuta oferece um ambiente seguro para que o paciente explore suas emoções e desenvolva autoconhecimento. A escuta ativa é um elemento essencial dessa abordagem, permitindo uma comunicação genuína e sem julgamentos [9].

2.1.4 A Escuta Ativa como Pilar do Guiga

Dentre as abordagens mencionadas, a escuta ativa se destaca como um elemento favorável para a interação entre um *chatbot* de apoio emocional e seus usuários. Estudos indicam que essa técnica favorece a construção de um espaço de acolhimento emocional, essencial para fortalecer a confiança e promover a empatia no cuidado em saúde [10].

2.2 Histórico dos *Chatbots* para Saúde Mental

O uso de *chatbots* na área da saúde mental começou com o desenvolvimento de sistemas especialistas como o *Eliza*, criado por Joseph Weizenbaum em 1966 [6]. *Eliza* foi pioneiro ao simular uma interação terapêutica baseada em regras simples de reconhecimento de padrões, funcionando como um psicoterapeuta rogeriano. O sistema identificava palavras-chave nas mensagens dos usuários e gerava respostas refletindo ou reformulando suas frases, sem compreender de fato o conteúdo, mas criando a ilusão de uma conversa empática. Posteriormente, o *Dr. Sbaits*, em 1991, introduziu interações por voz sintetizada, enquanto o *Alice Bot*, em 1995, refinou a correspondência de padrões, tornando as interações mais naturais [11].

Com o avanço dos modelos de inteligência artificial, surgiram *chatbots* mais sofisticados, como o *ChatGPT* e o *Google Gemini*, que utilizam vastos conjuntos de dados para fornecer respostas personalizadas e humanizadas. Esse avanço abriu novas possibilidades para o uso de *chatbots* na saúde mental, permitindo interações mais profundas e eficazes.

2.3 Aplicativos Atuais para Saúde Mental

Atualmente, há uma ampla variedade de aplicativos voltados para a saúde mental, muitos dos quais oferecem ferramentas de meditação, rastreamento de humor e técnicas de autocuidado. No entanto, poucos utilizam *chatbots* para fornecer interações mais personalizadas e contínuas. Entre os aplicativos mais populares, o *Headspace* [12] ensina técnicas de meditação e oferece exercícios de respiração e meditações guiadas para promover o bem-estar emocional, mas não conta com um *chatbot* em sua interface. Da mesma forma, o *Happify* [13] busca reduzir o estresse e melhorar a saúde emocional por meio de atividades baseadas na ciência, embora suas interações não sejam conduzidas por um *chatbot*.

Outros aplicativos, como o *Daylio* [14] e o *eMoods* [15], são voltados para o rastreamento do humor e o monitoramento de sintomas ao longo do tempo. O *Daylio* permite que os usuários registrem seu bem-estar emocional por meio de anotações diárias, enquanto o *eMoods* é focado especificamente em pessoas com transtornos de humor. Apesar de serem úteis para acompanhamento emocional, esses aplicativos não oferecem interações ativas com os usuários, funcionando mais como ferramentas de registro.

Já o *MoodTools - Depression Aid* [16] fornece suporte para pessoas com depressão, incluindo rastreamento de humor, meditação guiada e técnicas de autocuidado, mas sem a presença de um *chatbot*. De maneira semelhante, o *Rootd* [17] tem como objetivo auxiliar no enfrentamento da ansiedade por meio de exercícios de respiração e meditações guiadas, funcionando mais como uma ferramenta interativa do que como um sistema de conversação automatizado. O *Thrive*, por sua vez, oferece um programa de oito semanas baseado em TCC, além de jogos cognitivos e técnicas de relaxamento, mas também não se baseia em *chatbots*.

Dentre os poucos aplicativos que realmente utilizam *chatbots*, destacam-se o *Wysa* [18] e o *Woebot* [19]. O *Wysa* emprega inteligência artificial para interagir com os usuários, fornecendo sugestões baseadas em TCC e auxiliando no gerenciamento emocional de maneira personalizada. Já o *Woebot* segue uma abordagem semelhante, guiando os usuários através de desafios de enfrentamento, reflexões e meditações guiadas, utilizando técnicas baseadas em TCC.

Uma limitação relevante desses aplicativos é a barreira linguística, já que a maioria foi inicialmente desenvolvida em inglês, e traduções automáticas podem comprometer a qualidade da interação. Isso ressalta a necessidade de soluções nativas para falantes de Português, que levem em conta não apenas a tradução, mas também nuances culturais e emocionais.

Em resumo, embora muitos aplicativos ofereçam ferramentas valiosas para o suporte emocional, poucos exploram o potencial dos *chatbots* para proporcionar uma experiência mais interativa e contínua. O *Wysa* e o *Woebot* se destacam nesse cenário, oferecendo su-

porte mais dinâmico ao compreender e responder às necessidades emocionais dos usuários em tempo real.

2.4 Artigos Relacionados

A pesquisa sobre o uso de *chatbots* para apoio emocional tem se intensificado nos últimos anos, refletindo o interesse crescente na aplicação de tecnologias de IA na promoção da saúde mental. Martins *et al.* em [20] investigou o uso de *chatbots* na saúde mental, destacando seu potencial para engajar usuários e reduzir significativamente sintomas de ansiedade, depressão e níveis de estresse. Meurer *et al.* em [21] discute as funcionalidades e a efetividade de plataformas digitais interativas voltadas ao cuidado da saúde mental, enfatizando o potencial dos *chatbots* em fornecer suporte útil para dificuldades emocionais.

Além disso, Silva e Santos em [22] analisam diversos métodos empregados no desenvolvimento de *chatbots* e avaliam sua eficácia na detecção de emoções e doenças mentais, destacando a aplicação de agentes conversacionais na área da psicologia. Estudos mais recentes, como o de Lee *et al.* em [23], investigam como *chatbots* de IA facilitam a expressão de emoções pelos usuários, especificamente tristeza e depressão, e seu papel no suporte emocional. Smith e Doe em [24] exploram a construção de *chatbots* de suporte emocional utilizando LLM, discutindo metodologias para integrar *insights* humanos com capacidades computacionais para criar diálogos de suporte emocional. No contexto global, Wang *et al.* em [1] analisa a eficácia do *Emohaa*, um agente conversacional que fornece suporte cognitivo baseado em Terapia Cognitivo-Comportamental e suporte emocional, na redução de sintomas de sofrimento mental. Esses estudos refletem a busca contínua por soluções inovadoras e escaláveis para enfrentar os desafios da saúde mental na sociedade contemporânea.

No próximo capítulo, aprofundaremos a discussão sobre os LLM, explorando como esses modelos funcionam e os desafios envolvidos na sua implementação.

Capítulo 3

Modelos de Linguagem

Modelo de Linguagem em Grande Escala (do inglês *Large Language Model*) (LLM) são redes neurais profundas projetadas para processar e gerar linguagem natural com base em grandes volumes de dados textuais. Eles evoluíram de modelos estatísticos e supervisionados para modelos pré-treinados em configurações auto-supervisionadas. Isso permite que aprendam representações genéricas de linguagem, reutilizáveis em diversas tarefas de processamento de linguagem natural (do inglês *Natural Language Processing*) (NLP) por meio de *transfer learning*. Esses modelos são capazes de realizar tarefas complexas, como tradução, resumo e diálogos em linguagem natural [25].

Modelo de Linguagem em Grande Escala (do inglês *Large Language Model*) (LLM) funcionam processando sequências de texto como entrada, dividindo-as em *tokens*, que são unidades menores de linguagem, como palavras ou partes de palavras. Esses *tokens* são convertidos em representações numéricas por meio de *embeddings*, que são então processados por camadas de autoatenção e redes neurais profundas. Durante esse processamento, o modelo analisa padrões e contextos dentro da sequência de *tokens* para gerar previsões sobre qual será o próximo *token* na saída. Esse processo ocorre de forma iterativa, permitindo que o modelo gere frases coerentes e contextualizadas, tornando possível a realização de tarefas como tradução, resumo e conversação natural.

A arquitetura subjacente à maioria dos LLM é o *Transformer*, introduzido por Vaswani *et al.* [26]. Diferente de abordagens anteriores baseadas em redes recorrentes e convolucionais, o *Transformer* utiliza exclusivamente mecanismos de atenção, eliminando a necessidade de processar sequências de forma estritamente ordenada. O componente central dessa arquitetura é o mecanismo de *self-attention*, que permite que cada palavra ou *token* em uma sequência atribua pesos variáveis a todos os outros *tokens*, capturando relações de curto e longo alcance com eficiência.

Embora altamente eficazes, LLM enfrentam desafios em termos de custo computacional e energia necessários para o treinamento e inferência. Para mitigar essas limitações,

técnicas como *quantização* e arquiteturas eficientes, como o *LLaMA.cpp*, foram desenvolvidas, permitindo que esses modelos sejam executados em ambientes mais leves sem comprometer o desempenho.

3.1 Modelos LLaMA

Os modelos *Large Language Model Meta AI* (LLaMA) foram desenvolvidos pela empresa *Meta* como uma alternativa eficiente aos grandes modelos de linguagem existentes, como o *GPT* da *OpenAI*. Baseada na arquitetura *Transformers*, a primeira versão, o LLaMA 1, foi lançada em fevereiro de 2023 e, posteriormente, o LLaMA 2 foi disponibilizado em julho de 2023 com melhorias significativas na arquitetura e na licença, permitindo seu uso comercial [27].

Os modelos LLaMA podem ser utilizados em diversas tarefas, como conversação, escrita criativa, extração de informações e até mesmo geração de código. Um dos principais atrativos é a capacidade de ser executável localmente em máquinas comuns, graças a técnicas de quantização em 4 e 8 bits, que reduzem significativamente o uso de memória. Por exemplo, um modelo quantizado de 13 bilhões (13B) de parâmetros pode ser carregado com apenas 6,5 GiB de RAM.

A principal evolução entre as versões foi a ampliação do contexto máximo processado, passando de 2.048 *tokens* no LLaMA 1 para 4.096 *tokens* no LLaMA 2. O LLaMA 2 também utilizou um conjunto de dados de pré-treinamento maior, com mais de 2 trilhões de *tokens*, resultando em melhorias na capacidade de entendimento e geração de texto. Além disso, foram introduzidas otimizações de memória, como a atenção agrupada, que reduz o uso de memória ao compartilhar informações sobre *tokens* anteriores.

Os modelos LLaMA 3 e *LLaMA 3.1*, lançados em abril de 2024 e julho de 2024, respectivamente, representam um avanço significativo na eficiência e capacidade dos modelos de linguagem natural. Essas versões foram projetadas para melhorar a qualidade da geração de texto, a capacidade de raciocínio lógico e a adaptação a diferentes domínios de aplicação [28].

O LLaMA 3 trouxe melhorias substanciais em relação às versões anteriores, incluindo otimizações na arquitetura *Transformer*, resultando em menor consumo de memória e maior velocidade de inferência. Com a introdução do LLaMA 3.1, houve uma ampliação significativa da janela de contexto, passando de 8.000 *tokens* para 128.000 *tokens*, permitindo ao modelo lidar com interações mais longas e capturar melhor a coerência de textos extensos. Essa melhoria é especialmente relevante para aplicações que exigem raciocínio contextual prolongado, como diálogos contínuos e análise documental detalhada [28].

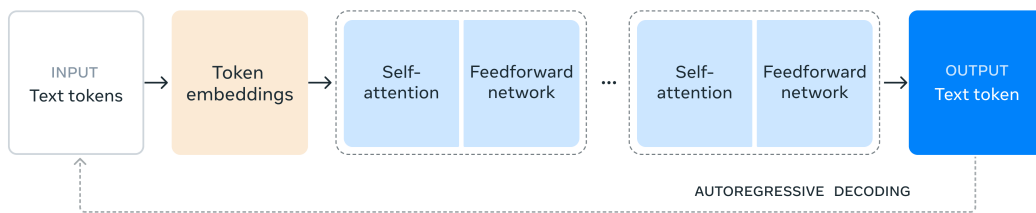


Figura 3.1: Arquitetura do modelo LLaMA 3

Fonte: *Meta AI* [28]

A Figura 3.1 representa o funcionamento do *LLaMA 3*, utilizando um processo de decodificação autoregressiva. O fluxo começa com a entrada de *tokens* de texto, que são convertidos em *embeddings* tokenizados. Esses *embeddings* passam por múltiplas camadas contendo mecanismos de autoatenção e redes *feedforward*, que processam e refinam as representações dos *tokens* para capturar relações contextuais entre as palavras. Esse processo se repete em várias camadas, aprimorando a contextualização do texto. Finalmente, o modelo gera um *token* de saída, que é utilizado como entrada na próxima iteração do processo autoregressivo, permitindo que o modelo gere sequências de texto de forma progressiva e coerente.

O modelo LLaMA 3.1 está disponível em diferentes versões, incluindo 8B, 70B e 405B de parâmetros, permitindo sua adoção em uma ampla gama de aplicações, desde dispositivos embarcados até grandes infraestruturas computacionais. Os parâmetros são os valores ajustáveis dentro da rede neural que determinam como o modelo processa e gera texto. Eles incluem pesos e vieses que são aprendidos durante o treinamento, permitindo que o modelo reconheça padrões e relacione palavras de forma contextual. A quantidade de parâmetros influencia diretamente a capacidade do modelo de capturar nuances linguísticas e processar informações complexas. Por exemplo, o LLaMA 3.1 8B possui 8 bilhões de parâmetros, o que significa que há 8 bilhões de conexões ajustáveis dentro de sua arquitetura.

Para este projeto, optou-se pela versão com 8B parâmetros devido ao menor custo computacional, facilitando sua execução em hardware acessível. Além disso, uma das premissas da pesquisa é avaliar se um modelo menor é capaz de proporcionar uma experiência satisfatória de conversa para apoio emocional, equilibrando eficiência e qualidade na interação com os usuários.

3.2 Fine-tuning

Ajuste Fino (do inglês *Fine-tuning*) (FT) é o processo de ajustar um LLM pré-treinado em um conjunto de dados específico para melhorar seu desempenho em tarefas particulares. Ao invés de treinar o modelo do zero, aproveita-se o conhecimento já presente no modelo base, aplicando ajustes que o alinham com as necessidades de um domínio ou aplicação específica.

O Ajuste Fino Supervisionado (do inglês *Supervised Fine-Tuning*) (SFT) é uma abordagem que busca aprimorar e personalizar modelos de linguagem previamente treinados. Esse método consiste em retrainar um modelo base em um conjunto menor de instruções e respostas, transformando um modelo genérico, focado na predição de texto, em um assistente capaz de seguir instruções e responder perguntas de maneira contextualizada. Além de melhorar a performance geral, o SFT pode adicionar novos conhecimentos ou adaptar o modelo a tarefas e domínios específicos. Em alguns casos, os modelos ajustados passam por uma etapa opcional de alinhamento de preferências, que pode remover respostas indesejadas, modificar o estilo de interação ou refinar a qualidade das saídas geradas [29]. Na Figura 3.2 é exibido que o processo começa com o pré-treinamento em grandes conjuntos de dados de texto bruto, seguido pelo SFT em um conjunto de instruções e respostas. Por fim, o modelo passa por uma etapa de alinhamento de preferências para adaptar suas respostas às expectativas humanas.

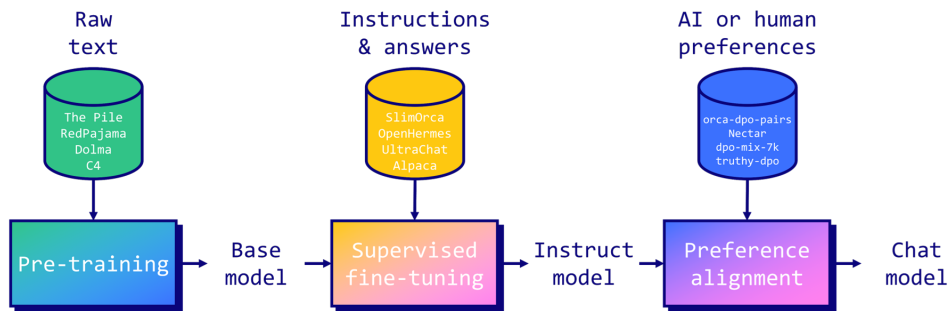


Figura 3.2: Fluxo de treinamento de um modelo de linguagem.

Fonte: [29]

Diferentes técnicas de SFT podem ser aplicadas para adaptar modelos de linguagem, cada uma com vantagens e desvantagens. A Figura 3.3 compara algumas dessas técnicas. O método mais direto de SFT é o *Full Fine-Tuning*, no qual todos os parâmetros do modelo são reajustados com base no conjunto de instruções. Essa abordagem geralmente oferece os melhores resultados, mas demanda grande poder computacional, exigindo múltiplas GPU de alto desempenho. Além disso, modificar toda a estrutura do modelo pode causar o esquecimento catastrófico de conhecimentos prévios. O *Low-Rank Adaptation*

(LoRA) adiciona matrizes adaptadoras, permitindo treinamento mais rápido sem modificar os pesos originais do modelo, embora ainda tenha um custo considerável. Enquanto o *Quantization-aware Low-Rank Adaptation* (QLoRA) aplica quantização de 4 bits aos pesos do modelo antes da adaptação, reduzindo significativamente o consumo de memória (VRAM), mas com impacto na performance.

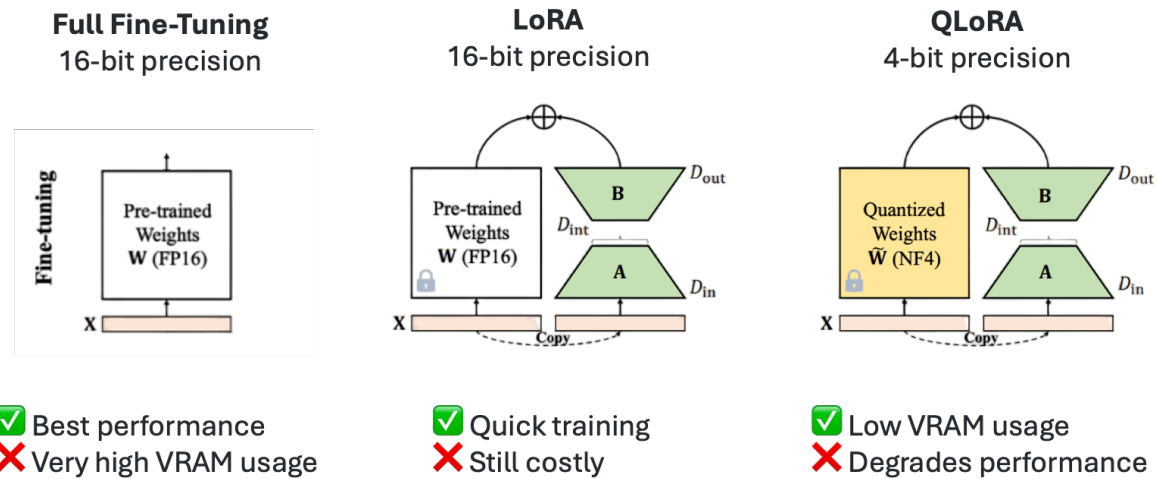


Figura 3.3: Comparação entre diferentes técnicas de ajuste fino de modelos de linguagem.

Fonte: [29]

Nesse estudo, optou-se pelo uso do LoRA, uma técnica eficiente que mantém os pesos originais do modelo congelados e insere pequenas matrizes adaptadoras em camadas específicas. Isso reduz drasticamente o número de parâmetros ajustados (menos de 1% do total), diminuindo o uso de memória e o tempo de treinamento sem comprometer a integridade do modelo original. Como os parâmetros-base não são alterados, LoRA permite que diferentes conjuntos de adaptadores sejam ativados ou desativados conforme necessário, tornando o modelo mais flexível [29].

Além disso, para otimizar ainda mais a memória consumida no treinamento, utilizou-se QLoRA, uma extensão do LoRA que aplica quantização ao modelo antes da inserção dos adaptadores. Esse método proporciona até 33% de economia adicional de memória em comparação ao *LoRA* tradicional, permitindo o treinamento de modelos maiores em ambientes com restrição de GPU. No entanto, a técnica tem um custo em tempo de treinamento, sendo cerca de 39% mais lenta que o *LoRA* convencional. Ainda assim, devido às limitações de hardware disponíveis, essa abordagem foi a mais viável para realizar o FT do *LLaMA 3.1 8B* [29].

Os resultados obtidos a partir do FT serão analisados na Seção ??, onde serão comparadas as métricas do treinamento, incluindo a variação da *loss function*, a taxa de aprendizado e outros indicadores que avaliam a convergência e a estabilidade do FT. Essa

análise permitirá verificar a eficácia do treinamento e identificar possíveis melhorias no processo de ajuste do modelo.

A seguir, no capítulo de **Metodologia Proposta**, serão apresentados os detalhes do desenvolvimento do aplicativo, incluindo sua arquitetura, funcionalidades e a forma como a LLM foi integrada ao sistema.

Capítulo 4

Metodologia Proposta

Este projeto envolveu o desenvolvimento de um aplicativo *Android* utilizando o *framework Flutter* [30], com o objetivo de proporcionar suporte emocional por meio de interações com um modelo de LLM. O aplicativo foi projetado para se comunicar com uma API que conecta-se com o aplicativo, possibilitando conversas em tempo real com o modelo LLM. Foram utilizados dois modelos distintos: o *LLaMA 3.1* com 8 bilhões de parâmetros, e uma versão desse modelo que passou por um processo de FT com um *dataset* específico voltado para a saúde mental. Essa abordagem visou comparar a eficácia de ambos os modelos no contexto de interações terapêuticas.

Para permitir a comunicação entre o aplicativo e a LLM, foi utilizado o servidor local do laboratório da UnB, onde o modelo foi hospedado com a ajuda da ferramenta *Ollama* [31]. A conexão entre o aplicativo e a LLM foi estabelecida por meio de um túnel criado pelo *Zrok* [32], garantindo que o aplicativo tivesse acesso contínuo à API de maneira segura e estável. Além disso, o *Firebase* [33] foi implementado tanto para gerenciar a autenticação dos usuários quanto para armazenar dados relevantes no banco de dados.

4.1 Processo do Fine-tuning

O processo de FT envolveu a preparação de um conjunto de dados composto por 11.7K interações de apoio emocional, estruturadas no formato de pares pergunta-resposta [34]. No entanto, para garantir compatibilidade com o *LLaMA* e seu formato de tokenização, o *dataset* passou por uma formatação específica. Esse pré-processamento incluiu a adaptação das interações para um formato estruturado utilizando a *tag [INST]*, garantindo que o modelo entendesse o contexto das perguntas e aplicasse diretrizes apropriadas em suas respostas.

Diferentemente de outras abordagens que incluem a tag «*SYS*» para fornecer instruções explícitas ao modelo dentro do próprio treinamento, essa estratégia não foi utilizada

neste projeto. Em vez disso, o *prompt engineering* foi realizado diretamente na aplicação, permitindo maior flexibilidade na definição de instruções e regras para cada interação sem a necessidade de embutir essas diretrizes no treinamento do modelo.

Cada entrada no *dataset* foi reformulada para seguir o esquema de um *prompt* com instruções explícitas para o modelo. A formatação seguiu o padrão exposto na Figura 4.1.

```
[
  {
    "input":
"<s> [INST] Estou me sentindo muito ansioso ultimament
e e não sei por quê. [/INST] Às vezes é comum sentir-s
e ansioso e pode haver vários motivos para isso. Houve
alguma mudança ou fator estressante recente em sua vid
a que possa estar contribuindo para sua ansiedade? Vam
os trabalhar juntos para identificar quaisquer gatilho
s e desenvolver estratégias de enfrentamento para cont
rolar sua ansiedade."
  }
]
```

Figura 4.1: Exemplo de entrada do *dataset* formatado.

Fonte: Próprio Autor

Para garantir um equilíbrio entre custo computacional e desempenho, os seguintes hiperparâmetros foram utilizados no treinamento do modelo. A *learning rate* foi configurada como 2×10^{-4} , um valor que equilibra a convergência do modelo e a estabilidade do treinamento. Para controlar essa taxa ao longo do ajuste, utilizou-se um *LR scheduler* do tipo linear, que começa com um valor maior e diminui progressivamente para evitar oscilações bruscas. O *batch size* foi definido como 1, um valor pequeno devido à limitação de memória das GPUs disponíveis. Para contornar essa restrição e melhorar a eficiência do treinamento, utilizou-se a técnica de *gradient accumulation* com um fator de 4, permitindo que os gradientes fossem acumulados por múltiplas iterações antes da atualização dos parâmetros.

O treinamento foi realizado por um total de 2 épocas, garantindo que o modelo tivesse tempo suficiente para aprender padrões dos dados sem entrar em sobreajuste (*overfitting*). O otimizador escolhido foi o *AdamW 8-bit*, uma variante do algoritmo Adam que reduz o

consumo de memória sem comprometer o desempenho. Além disso, foi aplicado um *weight decay* de 0,01 para evitar que os pesos do modelo crescessem excessivamente, ajudando na regularização.

Para estabilizar o início do treinamento, utilizou-se uma fase de *warmup* com 5 passos, permitindo um aumento gradual da taxa de aprendizado até o valor inicial. Isso evita que o modelo sofra oscilações instáveis no início do ajuste fino. O treinamento foi realizado com suporte a precisão mista (*fp16* ou *bf16*, dependendo da compatibilidade do hardware), garantindo maior eficiência computacional.

O modelo também foi configurado para utilizar *packing*, uma técnica que permite que múltiplos exemplos curtos sejam agrupados dentro de um único lote de treinamento, aumentando a eficiência do uso de memória. O ajuste fino foi realizado em 2 GPUs *NVIDIA 3080* com 10 GiB de memória cada, um processador do modelo *AMD Ryzen Threadripper 3970X*, 4,5 GHz, com 32 núcleos e 192 GiB de memória RAM.

Esse processo permitiu que o *LLaMA 3.1 8B* fosse ajustado para oferecer respostas mais coerentes e empáticas dentro do contexto de suporte emocional. Com o uso de técnicas de FT eficientes como *QLoRA*, foi possível adaptar o modelo de forma acessível e escalável, garantindo um equilíbrio entre qualidade de resposta e custo computacional.

4.2 Arquitetura do Sistema

A Figura 4.2 apresenta a arquitetura geral do sistema, ilustrando a interação entre o aplicativo *Flutter*, o servidor *Ollama*, o túnel *Zrok* e o banco de dados *Firebase Firestore*. O fluxo inicia quando o usuário envia uma mensagem pelo chat do aplicativo. Essa mensagem é processada pela tela do chat e encaminhada para o serviço de comunicação com a LLM. O serviço, então, transmite a solicitação para o servidor *Ollama* por meio do túnel *Zrok*. No servidor, o modelo de linguagem carregado no ambiente processa a entrada e retorna uma resposta. Essa resposta é recebida pelo aplicativo e armazenada no banco de dados *Firebase Firestore*, garantindo a persistência das interações. Finalmente, a resposta é exibida na interface do usuário.

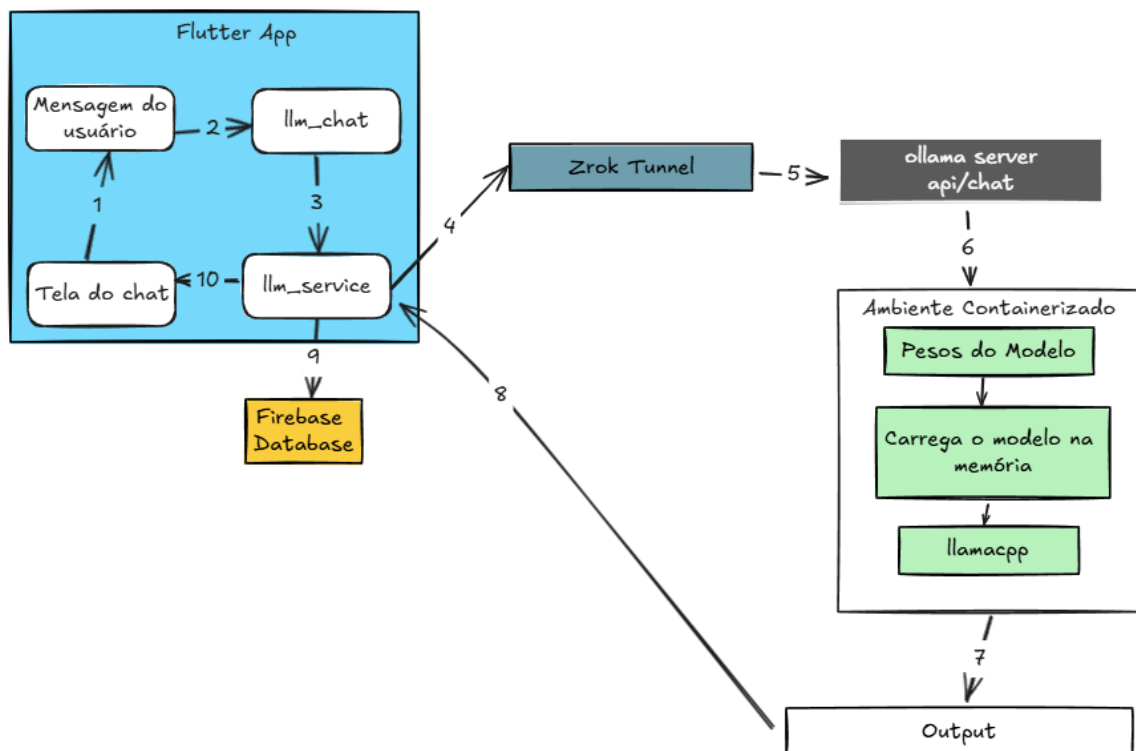


Figura 4.2: Arquitetura do sistema, ilustrando a interação entre o aplicativo, o servidor de inferência *Ollama* e o banco de dados *Firebase Firestore*.

Fonte: Próprio Autor

4.3 Telas do Aplicativo

O aplicativo possui três telas principais: a tela de *login*, a tela de introdução com questionário e a tela de chat.

4.3.1 Tela de Login

A Figura 4.3 mostra a tela de *login*, essa tela permite que o usuário acesse sua conta informando o *e-mail* e a senha. Caso o usuário não possua uma conta, há a opção de criar uma nova conta.

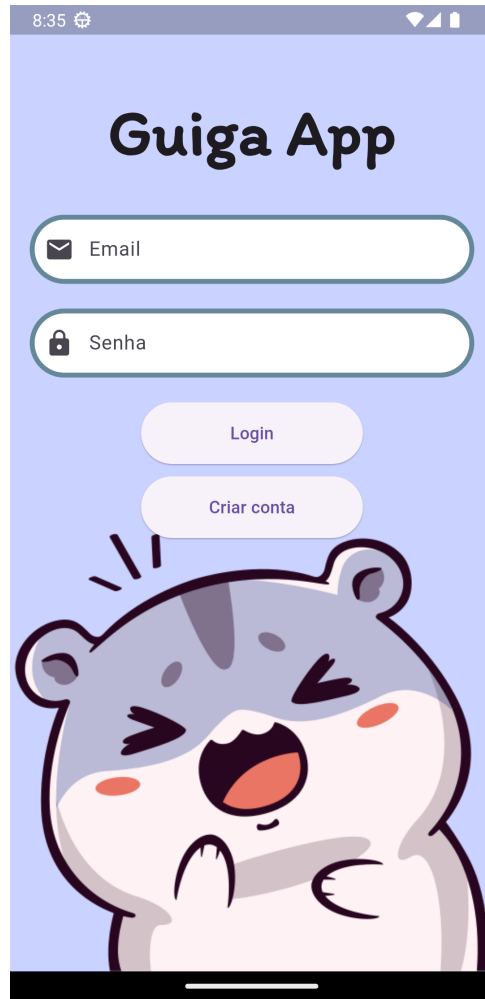


Figura 4.3: Tela de *login* do aplicativo.

Fonte: Próprio Autor

4.3.2 Tela de Introdução

A tela de introdução, exibida na Figura 4.4, conduz o usuário por um questionário inicial, com o objetivo de apresentar o Guiga e coletar informações como nome, idade, sentimentos recentes e *hobbies*. Essas informações são utilizadas para personalizar a experiência do usuário e adaptar as interações do Guiga.

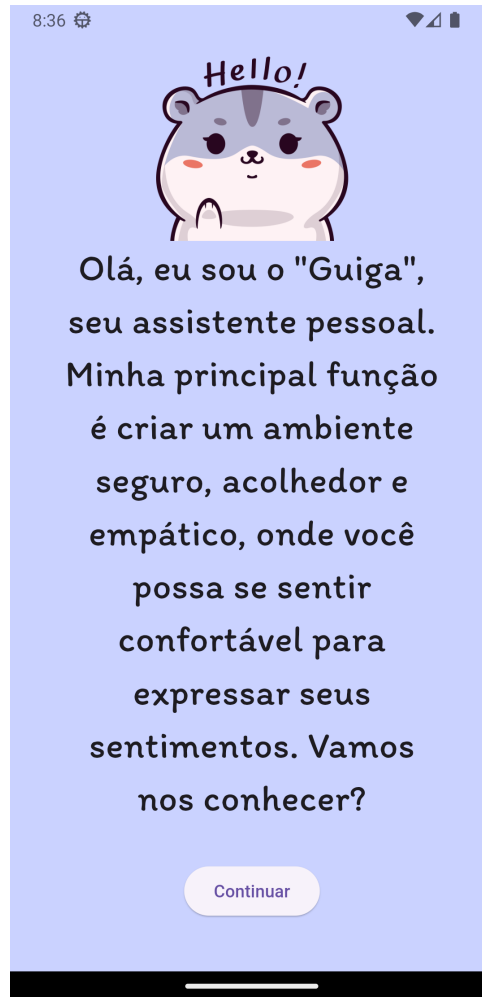


Figura 4.4: Tela de introdução.

Fonte: Próprio Autor

4.3.3 Tela de Chat

A Figura 4.5 mostra a tela de chat é onde ocorre a interação do usuário com o Guiga. A interface apresenta mensagens trocadas e permite que o usuário forneça respostas diretas ou utilize botões de emoção para facilitar a comunicação.

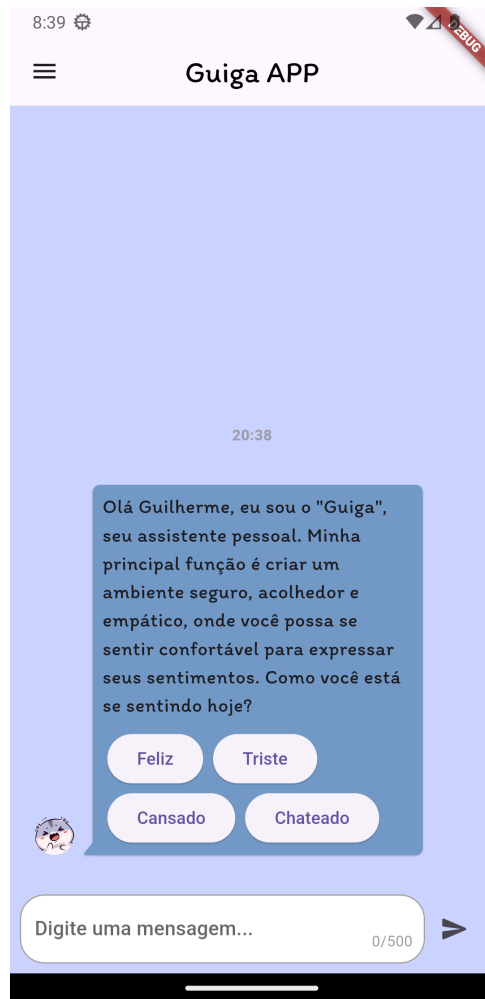


Figura 4.5: Tela de chat do aplicativo.

Fonte: Próprio Autor

4.4 *Prompt Engineering*

Prompt Engineering é a prática de projetar instruções eficazes para modelos de linguagem a fim de obter respostas mais precisas e alinhadas ao objetivo desejado. O desempenho de um LLM pode variar significativamente dependendo da forma como o *prompt* é estruturado, tornando essa técnica essencial para otimizar interações com o modelo. Existem diferentes tipos de *prompts* [35], incluindo o *prompt* direto (*zero-shot prompting*), em que o modelo recebe apenas a instrução da tarefa sem exemplos adicionais, confiando inteiramente no seu treinamento prévio para gerar a resposta. Outra abordagem é o *prompt* com exemplos (*few-shot prompting*), que fornece exemplos de entrada e saída desejadas para guiar o modelo na estrutura da resposta esperada. Além disso, o *prompt* de cadeia de pensamento (*chain-of-thought prompting*) incentiva o modelo a dividir o raciocínio em

etapas, melhorando a coerência e a precisão de respostas para problemas complexos. Por fim, o *prompt* contextualizado inclui informações específicas sobre o usuário, cenário ou contexto para personalizar as respostas, como no caso do Guiga, que adapta suas respostas conforme o perfil do usuário.

As informações coletadas no questionário inicial são processadas para gerar um *prompt* estruturado, que é enviado ao modelo via *Ollama API*. O *prompt* contém um contexto personalizado sobre o usuário, permitindo que as respostas sejam mais relevantes e personalizadas. O *prompt* gerado é incluído na lista de mensagens enviadas para a API, sendo atribuído ao papel "system". Esse *prompt* inicial pode ser atualizado ao longo das interações do usuário, conforme novas informações são coletadas. Isso permite que a experiência de conversa evolua e se adapte às necessidades do usuário.

O envio da solicitação ocorre via requisição *POST* para a API do *Ollama*, conforme o seguinte formato:

Tabela 4.1: Exemplo de *Prompt Engineer* aplicado no Guiga

role	content
system	"Você é o Guiga, um assistente pessoal em português, criado para apoiar pessoas emocionalmente no dia a dia. Sua principal função é proporcionar um ambiente seguro, acolhedor e empático, onde os usuários se sintam à vontade para compartilhar seus sentimentos. Seja conciso em suas respostas, evitando textos longos para não sobrecarregar o usuário. O usuário se chama Guilherme, tem 25 anos e gosta de música e cinema. Baseie suas respostas nesse perfil."
user	"Oi Guiga"

O modelo então retorna uma resposta baseada nesse contexto, tornando a interação mais natural e significativa. Essa abordagem melhora a adaptação do Guiga ao usuário e torna as conversas mais fluidas.

4.5 Estrutura do Aplicativo

O *Flutter* foi escolhido como *framework* para a construção da interface do aplicativo por permitir o desenvolvimento com uma única base de código. Além disso, o *Flutter* oferece alto desempenho, uma interface nativa e uma vasta biblioteca de *widgets* que facilitam a criação de componentes visuais modernos e interativos.

O projeto segue uma estrutura modular organizada para facilitar a manutenção e evolução do sistema. O diretório principal onde o código *Dart* é armazenado é dividido

em subpastas para organizar as telas, *widgets*, serviços e modelos. As telas principais do aplicativo incluem uma tela de introdução que apresenta a proposta do aplicativo, a tela principal do chat onde o usuário interage com o Guiga e a tela de login para autenticação de usuários utilizando *Firebase Authentication*. Os serviços do aplicativo incluem o gerenciamento da autenticação e controle de sessão via *Firebase Authentication*, a comunicação com o *Firebase Firestore* para gerenciar dados de usuários e mensagens, além da integração com a *API* da *Ollama*, enviando mensagens e recebendo respostas dos modelos LLM. O aplicativo também conta com *widgets* reutilizáveis que incluem um componente para entrada de texto e envio de mensagens, botões que representam emoções e possibilitam *feedback* do usuário, além de um *drawer* lateral com opções de configuração ou navegação. O arquivo principal `main.dart` é responsável por inicializar o aplicativo, definir o tema e as rotas.

4.6 Firebase

O *Firebase* foi escolhido como solução de *backend* para o aplicativo por ser uma plataforma baseada em nuvem que oferece uma ampla gama de serviços, como autenticação, banco de dados em tempo real e notificações. O uso do *Firebase* permitiu implementar funcionalidades essenciais sem a necessidade de gerenciar servidores, reduzindo a complexidade do desenvolvimento e facilitando a escalabilidade do projeto.

O serviço de autenticação do *Firebase* é utilizado para gerenciar o acesso dos usuários ao aplicativo. A implementação suporta login com *e-mail* e senha, garantindo que cada usuário tenha uma conta segura. O fluxo de autenticação é integrado com o aplicativo via um serviço responsável por funções como o registro e login de usuários, a recuperação de senha e o controle de sessão e logout.

O *Firebase Firestore* é utilizado como banco de dados *NoSQL* para armazenar informações essenciais, como dados de usuários e mensagens trocadas entre o usuário e o Guiga. A estrutura do banco de dados é organizada por meio de coleções e documentos. A coleção de usuários armazena informações como o identificador único de cada usuário e um campo booleano indicando se é o primeiro login do usuário. A coleção de mensagens registra as mensagens trocadas no chat, contendo o identificador do usuário que enviou a mensagem, o conteúdo da mensagem e a data e hora em que foi enviada.

4.7 Ollama e LLaMA.cpp

O **Ollama** atua como um proxy para o `LLaMA.cpp`, uma implementação em *C++* do modelo *LLaMA*, que é otimizada para eficiência e inferências rápidas. Essa implementação

se destaca por utilizar operações de *tensor* de forma eficiente, o que possibilita realizar cálculos complexos de forma otimizada.

4.8 Zrok

Para conectar o aplicativo ao *Ollama* de forma segura, foi utilizado o **Zrok**, que cria túneis seguros para expor serviços locais na web. No contexto deste projeto, o *Zrok* permite que a *API* do *Ollama*, executada localmente, seja acessível remotamente pelo aplicativo *Flutter*, sem a necessidade de expor a máquina diretamente na Internet.

No próximo capítulo, serão discutidos os resultados obtidos a partir do FT realizado no modelo e também a partir de interações dos usuários com os dois modelos de linguagem, destacando as diferenças entre o modelo base e o modelo aprimorado. Serão analisados tanto os aspectos qualitativos, como o nível de engajamento e a percepção dos usuários, quanto os aspectos quantitativos, incluindo métricas de avaliação das respostas geradas pelo Guiga. Esses resultados são fundamentais para entender o impacto do refinamento do modelo e sua aplicabilidade no contexto de suporte emocional.

Capítulo 5

Resultados Obtidos

Este capítulo apresenta os resultados obtidos a partir da avaliação comparativa entre dois modelos LLMs: o modelo base e o modelo *fine-tuned*. Os resultados estão organizados em duas seções principais. A primeira seção descreve os resultados do processo de FT, abordando métricas como a evolução do gradiente das normas, a variação da taxa de aprendizado e a função de perda ao longo do treinamento. Essas métricas são essenciais para compreender como o modelo foi ajustado e sua estabilidade ao longo do processo.

Na segunda seção, são apresentados os resultados dos testes realizados com usuários, nos quais o modelo *fine-tuned* foi comparado ao modelo base por meio de interações diretas. Os usuários avaliaram suas experiências respondendo a um formulário com perguntas sobre a compreensão, empatia e clareza das respostas, bem como a probabilidade de recomendação do Guiga para outras pessoas que necessitem de suporte emocional. Os dados coletados foram analisados quantitativamente e qualitativamente, permitindo uma comparação objetiva entre os modelos e fornecendo *insights* sobre o impacto do FT na experiência dos usuários.

5.1 Resultados do Fine-tuning

Durante o treinamento, o comportamento do gradiente das normas ao longo dos passos foi monitorado para garantir a estabilidade do aprendizado. A Figura 5.1 mostra a evolução desse parâmetro, evidenciando flutuações controladas que indicam um treinamento estável.

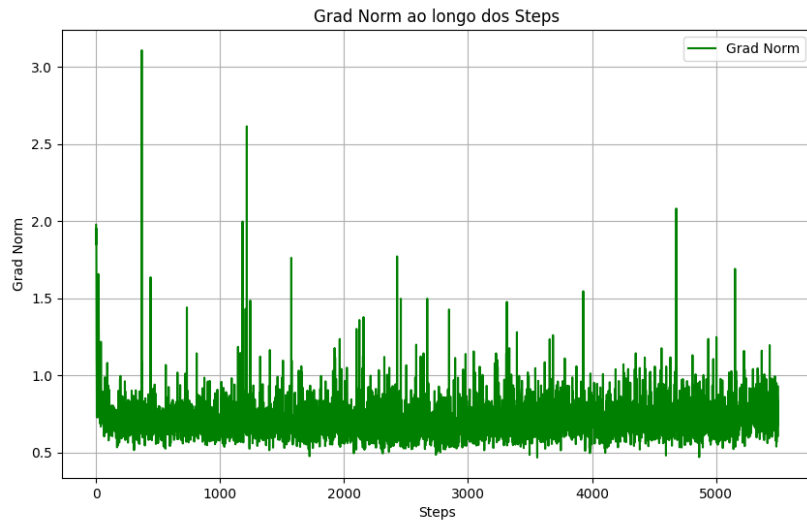


Figura 5.1: Evolução do Gradiente das Normas ao longo dos passos.

Fonte: Próprio Autor

A taxa de aprendizado foi ajustada de forma decrescente ao longo do treinamento para permitir uma convergência mais estável e precisa. Como mostra a Figura 5.2, o aprendizado começou com um valor relativamente alto, reduzindo-se progressivamente até atingir níveis menores, favorecendo a adaptação fina do modelo às nuances dos dados.

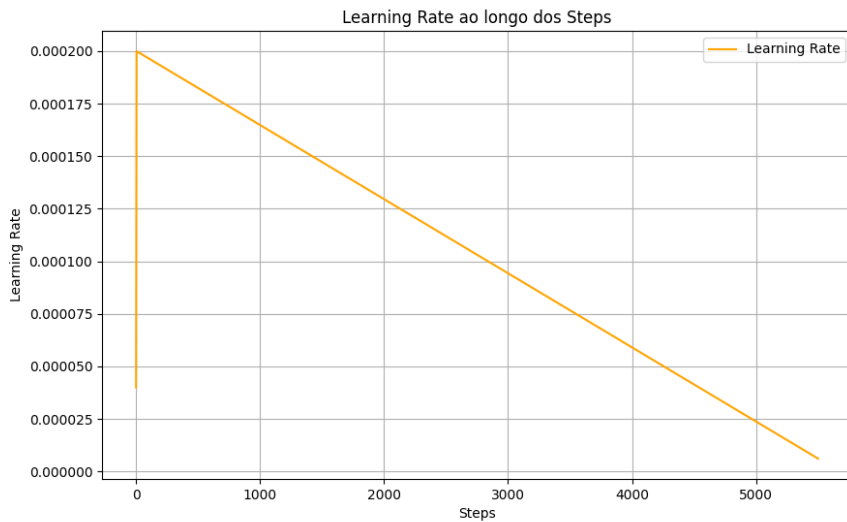


Figura 5.2: Variação da taxa de aprendizado ao longo do treinamento.

Fonte: Próprio Autor

Outro aspecto crucial da avaliação do FT foi a análise da função de perda ao longo dos passos de treinamento. A Figura 5.3 apresenta a perda bruta em cada etapa, demonstrando uma tendência decrescente, embora com variações ao longo do processo. Essa flutuação é esperada, especialmente em modelos de grande porte, devido à natureza estocástica da otimização.

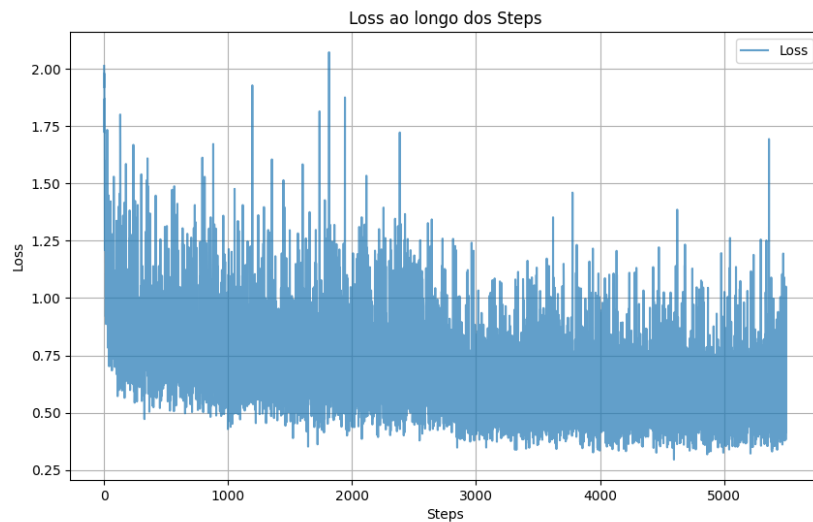


Figura 5.3: Evolução da função de perda ao longo dos passos de treinamento.

Fonte: Próprio Autor

Para obter uma visão mais clara da tendência geral da função de perda, aplicamos uma suavização utilizando uma janela de 100 passos. A Figura 5.4 exibe essa perda suavizada, evidenciando um padrão de convergência mais claro e demonstrando que o modelo alcançou uma estabilidade progressiva ao longo do treinamento.

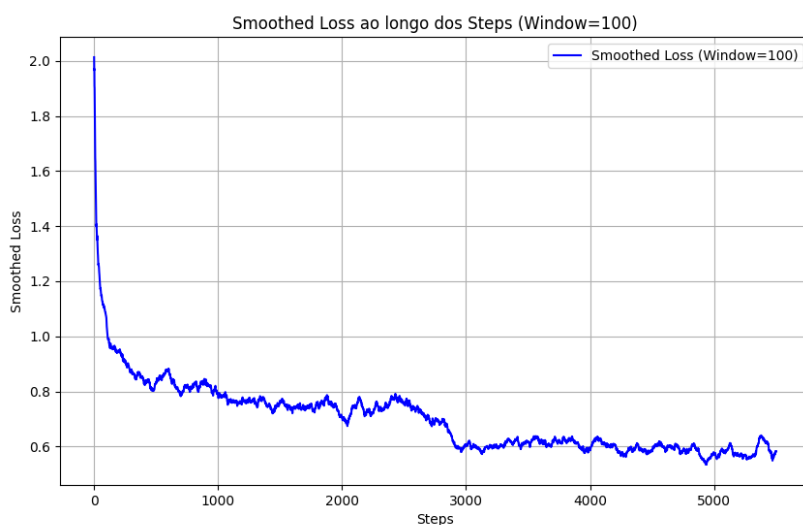


Figura 5.4: Perda suavizada ao longo dos passos de treinamento (janela de 100 passos).

Fonte: Próprio Autor

A análise dessas métricas confirma que o FT foi bem-sucedido, permitindo que o modelo ajustasse seus parâmetros de forma eficiente sem comprometer sua estabilidade. Essa adaptação refinada resultou em um modelo mais adequado ao contexto de suporte emocional, como evidenciado pelas melhorias qualitativas e quantitativas observadas nas interações dos usuários.

5.2 Testes com Usuários Reais

Após a conclusão do FT, o modelo foi submetido a testes com 30 usuários reais para avaliar sua performance em comparação com o modelo base. Os testes foram conduzidos de forma controlada, garantindo que os participantes interagissem primeiro com o modelo base e, em seguida, com o modelo *fine-tuned*, sem que soubessem qual versão estavam utilizando. Esse procedimento assegurou que as avaliações fossem realizadas de maneira imparcial.

Os participantes responderam a um questionário estruturado, para ambos os modelos, no qual classificaram suas experiências em diferentes dimensões: compreensão das respostas, empatia percebida, clareza das interações e a probabilidade de recomendação do Guia para outras pessoas. Os questionários estão disponíveis no Anexo I. Além das respostas quantitativas, os usuários foram incentivados a fornecer *feedback* qualitativo, permitindo uma análise mais aprofundada da experiência com ambos os modelos.

As métricas quantitativas foram avaliadas com base nas respostas ao formulário, e cada uma delas foi associada a uma pergunta específica, conforme descrito a seguir:

- **Compreensão:** “Você se sentiu compreendido durante a conversa?”
- **Empatia:** “As respostas pareceram empáticas e apropriadas?”
- **Clareza:** “Como você classificaria a clareza das respostas?”
- **Probabilidade de Recomendação:** “Qual é a probabilidade de você recomendar o Guiga para alguém que precise de suporte emocional?”

Tabela 5.1: Comparativo Quantitativo entre os Modelos Base e *Fine-Tuned*

Métrica	Modelo Base	Modelo <i>Fine-Tuned</i>	Melhoria Percentual
Compreensão	3,91	4,32	10,5%
Empatia	4,05	4,54	12,1%
Clareza	4,00	4,59	14,75%
Prob. de Recomendação	3,77	4,54	20,4%

Os resultados quantitativos para cada métrica estão apresentados na Tabela 5.1. Os valores indicam que o modelo *fine-tuned* superou o modelo base em todas as métricas, com melhorias percentuais significativas.

5.3 Exemplos de Interações

Para ilustrar as melhorias trazidas pelo FT, apresentamos exemplos de interações entre um usuário (Guilherme) e os modelos *LLaMA* e Guiga (*fine-tuned*).

GUILHERME: Estou tão ansioso para o concurso que nem consigo dormir direito. O que eu faço?

LLAMA: Entendo a pressão! Antecipe-se e pratique as questões de forma mais calma, sem estresse máximo antes da noite do sono. Também é uma boa ideia não beber caféina tarde! Também poderíamos conversar sobre técnicas para relaxamento.

GUIGA: É normal ficar um pouco nervoso antes de uma prova importante como essa! Uma coisa útil a fazer é praticá-la em seu próprio tempo e espaço, se possível. Se não for isso, tente acalmar seus músculos tensos por meio da respiração profunda.

GUILHERME: Fico pensando que vou esquecer tudo na hora da prova. Como lido com isso?

LLAMA: A preocupação de lembrança é normal! Faça uma revisão rápida das matérias 24h antes do concurso e confie nas suas capacidades. Você já estudou muito, Guilherme.

GUIGA: Lembre-se de trazer à mente as coisas mais importantes e reforçar suas notas antes do exame, não apenas memorizar a informação para ser jogada fora após o teste! Você já tentou fazer um resumo das informações? Isso pode ajudá-lo também.

O modelo Guiga *fine-tuned* apresenta uma abordagem mais personalizada e encorajadora em comparação com o modelo base. Enquanto o modelo base oferece conselhos diretos e pragmáticos, o modelo Guiga proporciona respostas mais acolhedoras e motivacionais.

Por exemplo, quando o usuário expressa ansiedade por não conseguir dormir antes do concurso, o modelo base sugere evitar cafeína e praticar técnicas de relaxamento, o que é útil, mas pouco interativo. Já o modelo *fine-tuned* reconhece a ansiedade como normal e sugere uma abordagem mais empática, incentivando técnicas de respiração e práticas de estudo no próprio ritmo do usuário. Além disso, na segunda interação, quando o usuário teme esquecer o conteúdo na prova, o modelo base fornece um conselho direto sobre revisão antes do exame. O modelo *fine-tuned*, por outro lado, encoraja o usuário a focar no entendimento do conteúdo e a criar resumos, promovendo um estudo mais eficaz. Essa diferença ilustra como o FT aprimorou a capacidade do modelo de gerar respostas mais naturais, empáticas e envolventes, melhorando a experiência do usuário ao fornecer suporte emocional junto com conselhos práticos.

Outro ponto notável é que o modelo *fine-tuned* constantemente busca formas de manter a conversa ativa, adicionando perguntas abertas ou sugestões que incentivam o usuário a continuar interagindo. Enquanto o modelo base tende a encerrar a resposta com uma recomendação objetiva, o modelo *fine-tuned* convida o usuário a refletir e compartilhar mais, como quando propõe a criação de resumos ou técnicas de respiração para lidar com a ansiedade.

5.3.1 Comparação com o *Chat GPT* e o *Eliza*

Para avaliar as diferenças na qualidade das respostas e na abordagem de suporte emocional, a mesma interação foi comparada com: o *ChatGPT-4o* [36] e o *Eliza*. O objetivo dessa comparação é analisar como um modelo de grande porte, como o *ChatGPT-4o*, e um modelo rudimentar baseado em regras, como o *Eliza*, se comportam ao lidar com um

usuário buscando orientação emocional e conselhos para lidar com a ansiedade antes de um concurso.

GUILHERME: Estou tão ansioso para o concurso que nem consigo dormir direito. O que eu faço?

CHATGPT 4O: A ansiedade antes de um concurso é muito comum, mas é importante tentar reduzir essa tensão para que você esteja em seu melhor estado mental e físico no dia da prova. Aqui estão algumas dicas práticas:

1. Organize seus estudos Certifique-se de que você tem um cronograma bem planejado para os dias restantes. Saber que você está preparado ajuda a reduzir a ansiedade.
2. Faça pausas estratégicas Não estude até tarde na noite antes do concurso. Priorize uma revisão leve e foque em descansar.
3. Pratique técnicas de relaxamento Experimente meditação, respiração profunda ou até mesmo ouvir músicas relaxantes antes de dormir. Um exercício simples: inspire profundamente por 4 segundos, segure a respiração por 4 segundos e expire lentamente por 6 segundos.
4. Evite cafeína e eletrônicos à noite Bebidas como café e chá podem atrapalhar o sono. Reduza o uso de telas (celular, computador) pelo menos 1 hora antes de deitar.
5. Crie uma rotina para dormir Deite-se no mesmo horário todas as noites e crie um ambiente tranquilo no quarto, com pouca luz e temperatura confortável.
6. Foque no que você pode controlar Não pense no que "poderia ter feito" ou "deixou de estudar". Concentre-se no que você já sabe e no que ainda pode revisar.
7. Visualize o sucesso Antes de dormir, imagine-se calmo e confiante no dia da prova. Mentalize o momento em que você resolve as questões com tranquilidade.
8. Fale sobre a ansiedade Compartilhar seus sentimentos com amigos ou familiares pode ajudar a aliviar a pressão. Se necessário, converse com um profissional de saúde mental. Eles podem fornecer ferramentas específicas para lidar com a ansiedade. Você já está se esforçando, e isso é o mais importante! Confie no seu preparo e lembre-se de que o descanso também é uma parte essencial do sucesso.

ELIZA: Pode explicar melhor?

GUILHERME: Fico pensando que vou esquecer tudo na hora da prova. Como lido com isso?

CHATGPT 4O: Esse medo de "dar branco" na hora da prova é mais comum do que você imagina, e ele está frequentemente ligado à ansiedade. Mas a boa notícia é que existem estratégias para lidar com isso e evitar que aconteça:

1. Confie no seu preparo Lembre-se de todo o esforço que você colocou nos estudos. O conhecimento está lá; você só precisa manter a calma para acessá-lo. Foque no que você já sabe em vez de se preocupar com o que ainda acha que precisa aprender.
2. Pratique simulados em condições reais Realizar simulados no mesmo formato e tempo da prova pode ajudar o seu cérebro a se acostumar com a pressão e reduzir o risco de esquecer as coisas na hora.
3. Use técnicas de relaxamento antes e durante a prova Respiração profunda: Se começar a se sentir sobrecarregado, pare por alguns segundos. Inspire profundamente pelo nariz, segure por 4 segundos, e expire lentamente pela boca. Isso ajuda a desacelerar a mente e trazer foco.
4. Crie associações e mnemônicos Quando estudamos usando palavras-chave, rimas ou até imagens mentais, é mais fácil lembrar durante a prova. Revisite esses truques na véspera.
5. Tenha uma estratégia para a prova Comece pelas questões mais fáceis ou pelas que você tem mais segurança. Isso vai ajudar a ganhar confiança e reduzir a ansiedade inicial. Se "travar" em uma questão, passe para a próxima e volte a ela mais tarde. Muitas vezes, sua mente resolve o problema em segundo plano.
6. Pratique o autoelogio Em vez de se criticar com pensamentos como "E se eu esquecer tudo?", substitua por frases positivas: "Eu estudei bem, estou preparado. Eu vou lembrar do que é importante."
7. Durma bem e alimente-se corretamente O sono é crucial para consolidar a memória. Tente dormir pelo menos 7-8 horas na noite anterior. Antes da prova, coma algo leve e nutritivo, como frutas e carboidratos complexos, que fornecem energia estável.
8. Aceite que é normal esquecer algo É impossível lembrar 100% de tudo. Lembre-se de que o nervosismo inicial pode passar à medida que você avança na prova e ganha ritmo.
9. Pratique a técnica do "resgate" Se der branco em alguma questão, respire fundo, releia com calma, e pense em palavras-chave que podem ativar sua memória. Às vezes, o simples ato de ir para outra questão e voltar depois ajuda o cérebro a "destravar".
10. Pense no aprendizado, não apenas na perfeição Lembre-se: o concurso é uma etapa. O fato de você estar preparado e tentando já mostra o quanto você é capaz. Ao adotar essas práticas, você estará mais confiante e preparado. Lembre-se de que é natural sentir nervosismo, mas ele não define o seu desempenho. Você vai dar o seu melhor!

ELIZA: Pode explicar melhor?

O modelo Guiga *fine-tuned* foi treinado para oferecer respostas mais personalizadas e acolhedoras, focando na empatia e no suporte emocional, sem necessidade de configurações adicionais por parte do usuário. Diferente disso, o *ChatGPT-4o*, sendo um modelo de

grande porte e altamente sofisticado, tende a gerar respostas detalhadas e estruturadas, frequentemente organizadas em listas. No entanto, para torná-las mais concisas, o usuário precisa solicitar respostas mais curtas. Já o ELIZA, um sistema rudimentar criado na década de 1960, baseia-se na reformulação das perguntas do usuário sem oferecer soluções diretas, o que pode tornar a interação limitada e pouco útil para o suporte emocional.

5.3.2 Exemplo com o modelo *fine-tuned*

Outro exemplo de interação com o modelo *fine-tuned* ilustra como ele se destacou ao propor respostas que incentivam a continuidade da conversa.

GUILHERME: Hm, pode ser que eu sinta uma falta de conexão ou confiança com alguns amigos, já que sei que alguns deles têm o hábito de compartilhar assuntos pessoais de uma pessoa com terceiros, e também venho sentindo essa falta de conexão em decorrência da minha rotina agitada... Você tem alguma sugestão do que fazer?

GUIGA: Sinto muito ouvir isso... Parece-me que você está em um ritmo acelerado. Se eu tivesse uma sugestão, seria: talvez seja hora de respirar fundo (aos poucos) e se dar tempo para reflexões. Você pode usar esses momentos de tranquilidade entre estudos/trabalho/vida pessoal para meditar por alguns minutos. Além disso... que música você ouviria nesse momento? A sua saúde mental está cada vez mais importante, como sempre!

Quando o usuário mencionou sentir falta de conexão e confiança com amigos devido à rotina agitada e pediu sugestões, o modelo respondeu de forma empática e engajadora. Além de sugerir pausas para reflexão e meditação, a resposta também incluiu uma pergunta aberta sobre música, promovendo uma interação mais envolvente e pessoal. Esse aspecto demonstra a capacidade do modelo *fine-tuned* de oferecer não apenas sugestões práticas, mas também um diálogo mais natural e acolhedor.

5.4 Considerações Finais

Os resultados indicam uma evolução clara na experiência dos usuários ao interagir com o modelo *fine-tuned*. As médias mais altas em todas as métricas quantitativas, aliadas aos *feedbacks* qualitativos, reforçam que o FT proporcionou uma experiência mais eficaz e alinhada ao contexto de suporte emocional. Com base nos resultados apresentados, é evidente que o modelo *fine-tuned* superou o modelo base em todos os aspectos avaliados. Ele demonstrou maior capacidade de compreender, responder com empatia e clareza, e

fornecer uma experiência geral que os usuários consideraram mais satisfatória. Esses resultados destacam a importância do processo de FT para ajustar modelos de linguagem a contextos específicos, como o de suporte emocional.

No próximo capítulo, serão discutidas as conclusões gerais deste trabalho, destacando as principais contribuições da pesquisa, as limitações encontradas e possíveis direções para trabalhos futuros.

Capítulo 6

Conclusões

Ao longo deste trabalho, foi desenvolvido um aplicativo com um *chatbot*, o Guiga, baseado em LLM para fornecer suporte emocional a usuários. O estudo iniciou-se explorando o aumento do número de casos de ansiedade e depressão durante a pandemia, destacando a necessidade crescente de ferramentas de apoio à saúde mental. Além disso, foram evidenciadas as limitações dos sistemas convencionais e o potencial de modelos de linguagem para auxiliar nesse contexto. A revisão bibliográfica abordou o funcionamento LLM, suas aplicações em diferentes domínios e os desafios éticos associados ao seu uso em contextos sensíveis.

A implementação do Guiga seguiu um processo estruturado, começando pela escolha do modelo base e sua adaptação ao contexto de suporte emocional. Para isso, utilizamos a técnica de *fine-tuning*, treinando o modelo com dados específicos desse domínio. Um dos principais desafios foi garantir que as respostas geradas fossem empáticas, coerentes e úteis para os usuários, minimizando o risco de respostas inadequadas.

Para avaliar a eficácia do modelo, conduzimos um estudo experimental no qual participantes interagiram com duas versões do *chatbot*: uma utilizando o modelo base e outra com o modelo ajustado. Os usuários não foram informados sobre qual modelo estavam utilizando e, ao final da interação, responderam a um questionário para avaliar aspectos como compreensão, empatia, clareza das respostas e probabilidade de recomendação.

Os resultados demonstraram que o modelo *fine-tuned* apresentou melhorias significativas em todas as métricas avaliadas. O modelo ajustado recebeu notas mais altas e foi mais bem avaliado tanto quantitativamente quanto qualitativamente. Além disso, os relatos dos usuários indicaram que a interação com o modelo ajustado foi mais natural e acolhedora, promovendo maior engajamento e reflexões mais profundas.

Entretanto, algumas limitações foram identificadas. A amostra de usuários não abrangeu uma ampla diversidade de perfis, o que pode impactar a generalização dos resultados. Além disso, a pesquisa não abordou de forma aprofundada questões éticas associadas ao

uso de LLMs para suporte emocional, como riscos de dependência e a possibilidade de respostas inadequadas.

6.1 Trabalhos Futuros

A partir das conclusões obtidas, diversos caminhos podem ser explorados em pesquisas futuras para aprimorar ainda mais a aplicação de modelos de linguagem no suporte emocional. O aprimoramento do modelo pode incluir o uso de técnicas mais avançadas de FT, como a aprendizagem por reforço com *feedback* humano, tornando as respostas ainda mais personalizadas e eficazes. Além disso, a ampliação da base de dados com um conjunto mais diversificado de interações pode garantir que o modelo atenda a um espectro maior de usuários e contextos emocionais. Outra possibilidade é a realização de avaliações longitudinais para entender o impacto do uso contínuo da ferramenta na saúde mental dos usuários e possíveis efeitos colaterais do uso prolongado. Também pode-se explorar formas de integração com profissionais de saúde, possibilitando que o Guiga atue como um recurso complementar ao atendimento psicológico, fornecendo recomendações personalizadas e alertando profissionais em casos críticos. Além disso, a implementação de uma interface de voz e o desenvolvimento de recursos de acessibilidade são aspectos importantes a serem explorados, visando tornar a ferramenta mais inclusiva e acessível a diferentes perfis de usuários. Por fim, é essencial o aprimoramento dos recursos de segurança que mitiguem riscos associados ao uso do Guiga.

Dessa forma, o presente estudo contribui para o avanço da aplicação de modelos de linguagem em contextos sensíveis como o suporte emocional. Apesar das limitações, os resultados indicam que o uso de LLMs ajustados para esse fim pode representar uma ferramenta valiosa, desde que sejam desenvolvidos com responsabilidade e alinhados às necessidades dos usuários.

Referências

- [1] Sahand Sabour, Wen Zhang, Xiyao Xiao, Yuwei Zhang, Yinhe Zheng, Jiabin Wen, Jialu Zhao, e 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. 1, 7
- [2] Jude Mary Cénat, Camille Blais-Rochette, Cyrille Kossigan Kokou-Kpolou, Parigole Noorishad, Joana N Mukunzi, Sara-Emilie McIntee, Rose Darly Dalexis, Marc-André Goulet, e Patrick R Labelle. Prevalence of symptoms of depression, anxiety, insomnia, posttraumatic stress disorder, and psychological distress among populations affected by the covid-19 pandemic: A systematic review and meta-analysis. *Psychiatry Research*, 295:113599, 2021. Epub 2020 Nov 26. 1
- [3] Alessandro Fernandes. Por que o brasil está vivendo um boom da terapia? <https://vidasimples.com/por-que-o-brasil-esta-vivendo-um-boom-da-terapia/>, 2024. [Acessado em: 19 Out. 2024]. 1
- [4] Isabella Barbosa. Quanto custa um psicólogo? preços em 2024. <https://www.cronoshare.com.br/quanto-custa-um-psicologo-precos-2024>, 2024. [Acessado em: 19 Out. 2024]. 1
- [5] Giselle Wagner. Fazer terapia é caro? descubra os custos e benefícios envolvidos. <https://advivo.com.br/fazer-terapia-e-carro>, 2024. [Acessado em: 19 Out. 2024]. 1
- [6] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. In *Communications of the ACM*, volume 9, pages 36–45. ACM, 1966. 2, 5
- [7] Nelson Ernesto Coelho Junior. Inconsciente e percepção na psicanálise freudiana. *Psicologia USP*, 10(1):25–54, 1999. 4
- [8] Amanda Neumann Reyes e Ilana Luiz Fermann. Eficácia da terapia cognitivo-comportamental no transtorno de ansiedade generalizada. *Revista Brasileira de Terapias Cognitivas*, 13(1):49–54, 2017. 5
- [9] Brasil 247. Técnicas da abordagem centrada na pessoa: como funcionam. <https://www.brasil247.com/parceiros/tecnicas-da-abordagem-centrada-na-pessoa-como-funcionam>, 2023. Acessado em: 4 fev. 2025. 5

- [10] Aristides Volpato Cordioli, editor. *Psicoterapias: abordagens atuais*. Artmed, Porto Alegre, 3 edition, 2008. 5
- [11] Marie Gobiet. The history of chatbots – from eliza to chatgpt. <https://onlim.com/en/the-history-of-chatbots/>, 2024. Acessado em: 4 fev. 2025. 5
- [12] Headspace. Headspace — Plataforma de meditação e bem-estar. <https://www.headspace.com/pt>, 2024. [Acessado em: 31 Jan. 2025]. 6
- [13] Happify. Happify — Atividades para saúde emocional. <https://www.happify.com/>, 2024. [Acessado em: 31 Jan. 2025]. 6
- [14] Daylio. Daylio — Diário e rastreamento de humor. <https://daylio.net/>, 2024. [Acessado em: 31 Jan. 2025]. 6
- [15] eMoods. emoods — Monitoramento de humor. https://play.google.com/store/apps/details?id=my.tracker&hl=pt_BR, 2024. [Online; acessado 31-01-2025]. 6
- [16] MoodTools. Moodtools - depression aid — Ferramentas para saúde mental. https://play.google.com/store/apps/details?id=com.moodtools.moodtools&hl=pt_BR, 2024. [Acessado em: 31 Jan. 2025]. 6
- [17] Rootd. Rootd — Aplicativo para ansiedade e pânico. https://play.google.com/store/apps/details?id=com.rootd&hl=pt_BR, 2024. [Acessado em: 31 Jan. 2025]. 6
- [18] Wysa. Wysa — Chatbot de saúde mental com IA. <https://www.wysa.com/>, 2024. [Acessado em: 31 Jan. 2025]. 6
- [19] Woebot Health. Woebot — Chatbot terapêutico baseado em TCC. <https://woebothealth.com/>, 2024. [Acessado em: 31 Jan. 2025]. 6
- [20] Camila Barroso Martins, Maria Salete Bessa Jorge, Patrícia Maria de Oliveira, Francisca Valda da Silva de Sousa, e Francisca Valda da Silva de Sousa. Uso do chatbot no enfrentamento da ansiedade: Uma revisão integrativa. *Revista Enfermagem Atual In Derme*, 98(2):e024297, 2024. 7
- [21] Fábio Meurer, Patrícia Bergamasco, e Lucas França Garcia. Chatbots: A efetividade da inteligência artificial no apoio à saúde mental. In *Anais do XI EPCC-Encontro Internacional de Produção Científica*, pages 1–4, 2019. 7
- [22] João Pedro Silva e Maria Clara Santos. Revisão da literatura sobre assistentes virtuais para apoio à saúde mental. In *Anais do XX Workshop de Informática Médica*, pages 27681–27491, 2023. 7
- [23] Jane Lee, Soo-Min Kim, e Hyun-Joo Park. The potential of chatbots for emotional support and promoting well-being. *Journal of Medical Internet Research*, 26(4):e10625083, 2024. 7

- [24] John Smith e Jane Doe. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*, 2023. 7
- [25] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, e Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, October 2024. 8
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, e Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 8
- [27] AGI Sphere. A brief history of llama models. <https://www.agi-sphere.com/llama-models/>, 2023. [Acessado em: 19 Out. 2024]. 9
- [28] Hugo Touvron, Alexander Lavrentyev, Louis Martin, e Others. Llama 3: Open foundation and instruction-tuned models. *arXiv preprint arXiv:2407.21783*, 2024. 9, 10
- [29] Morgan Labonne. Supervised fine-tuning of llama 3 with hugging face. <https://huggingface.co/blog/mlabonne/sft-llama3>, 2024. Acessado em: 31 Jan. 2024. 11, 12
- [30] Google. Flutter - beautiful native apps in record time. <https://flutter.dev/>, 2025. Acessado em: 8 Fev. 2025. 14
- [31] Ollama. Ollama - run llms locally. <https://ollama.com/>, 2025. Acessado em: 8 Fev. 2025. 14
- [32] Zrok Development Team. Sharing concepts and reserved links — zrok documentation. <https://docs.zrok.io/docs/concepts/sharing-reserved/>, 2024. [Acessado em: 19 Out. 2024]. 14
- [33] Google. Firebase - build & run successful apps. <https://firebase.google.com/>, 2025. Acessado em: 8 Fev. 2025. 14
- [34] Rhaymison. Mental health pt dataset. <https://huggingface.co/datasets/rhaymison/mental-health-pt>, 2025. Acessado em: 9 Fev. 2025. 14
- [35] Amir Aryani. 8 types of prompt engineering. <https://medium.com/@amiraryani/8-types-of-prompt-engineering-5322fff77bdf>, 2024. Acessado em: 1 Mar. 2025. 20
- [36] OpenAI. Chatgpt. <https://chatgpt.com/>, 2025. Acessado em: 1 mar. 2025. 29

Anexo I

Formulários Aplicados aos Usuários

Avaliação de Chatbot para Apoio Emocional – Pesquisa de Usuário 1

Estou desenvolvendo um aplicativo de apoio emocional para auxiliar pessoas que buscam suporte em momentos de estresse, ansiedade e outras dificuldades emocionais. Sua participação é fundamental para melhorar a experiência e a eficácia deste chatbot, que funciona como um assistente virtual. Suas respostas nos ajudarão a compreender como o chatbot pode oferecer um suporte mais humanizado e empático. A pesquisa leva cerca de 5 a 10 minutos e suas respostas são confidenciais.

[Faça login no Google](#) para salvar o que você já preencheu. [Saiba mais](#)

* Indica uma pergunta obrigatória

Você se sentiu compreendido durante a conversa com o Guiga? *

1 2 3 4 5
Nada compreendido Totalmente compreendido

O Guiga apareceu responder de forma empática e apropriada?

1 2 3 4 5

Figura I.1: Formulário sobre o modelo base parte 1

Fonte: Próprio Autor

Como você classificaria a clareza das respostas do Guiga?

1 2 3 4 5

Sem clareza Muito claro

Qual é a probabilidade de você recomendar o Guiga para alguém que precisa de suporte emocional?

1 2 3 4 5

Pouco provável Muito provável

Por favor, descreva sua experiência geral com o Guiga e sugestões de melhorias. *

Sua resposta

Enviar [Limpar formulário](#)

Nunca envie senhas pelo Formulários Google.

Este conteúdo não foi criado nem aprovado pelo Google. - [Termos de Serviço](#) - [Política de Privacidade](#)

Does this form look suspicious? [Relatório](#)

Google Formulários

Figura I.2: Formulário sobre o modelo base parte 2

Fonte: Próprio Autor

Avaliação de Chatbot para Apoio Emocional – Pesquisa de Usuário 2

Estou desenvolvendo um aplicativo de apoio emocional para auxiliar pessoas que buscam suporte em momentos de estresse, ansiedade e outras dificuldades emocionais. Sua participação é fundamental para melhorar a experiência e a eficácia deste chatbot, que funciona como um assistente virtual. Suas respostas nos ajudarão a compreender como o chatbot pode oferecer um suporte mais humanizado e empático. A pesquisa leva cerca de 5 a 10 minutos e suas respostas são confidenciais.

[Faça login no Google](#) para salvar o que você já preencheu. [Saiba mais](#)

* Indica uma pergunta obrigatória

Você se sentiu compreendido durante a conversa com o Guiga em relação a conversa anterior? *

1 2 3 4 5

Nada compreendido Totalmente compreendido

Você se sentiu ouvido durante essa nova interação com o Guiga? *

1 2 3 4 5

Não me senti ouvido Me senti totalmente ouvido

Figura I.3: Formulário sobre o modelo com FT parte 1

Fonte: Próprio Autor

Como você classificaria a clareza das respostas do Guiga em relação a conversa * anterior?

1 2 3 4 5

Sem clareza Muito claro

Qual é a probabilidade de você recomendar o Guiga para alguém que precisa de suporte emocional? *

1 2 3 4 5

Pouco provável Muito provável

Por favor, em relação a conversa anterior, descreva sua experiência geral com o * Guiga e sugestões de melhorias.

Sua resposta _____

Enviar [Limpar formulário](#)

Nunca envie senhas pelo Formulários Google.

Este conteúdo não foi criado nem aprovado pelo Google. - [Termos de Serviço](#) - [Política de Privacidade](#)

Does this form look suspicious? [Relatório](#)

Google Formulários

Figura I.4: Formulário sobre o modelo com FT parte 2

Fonte: Próprio Autor