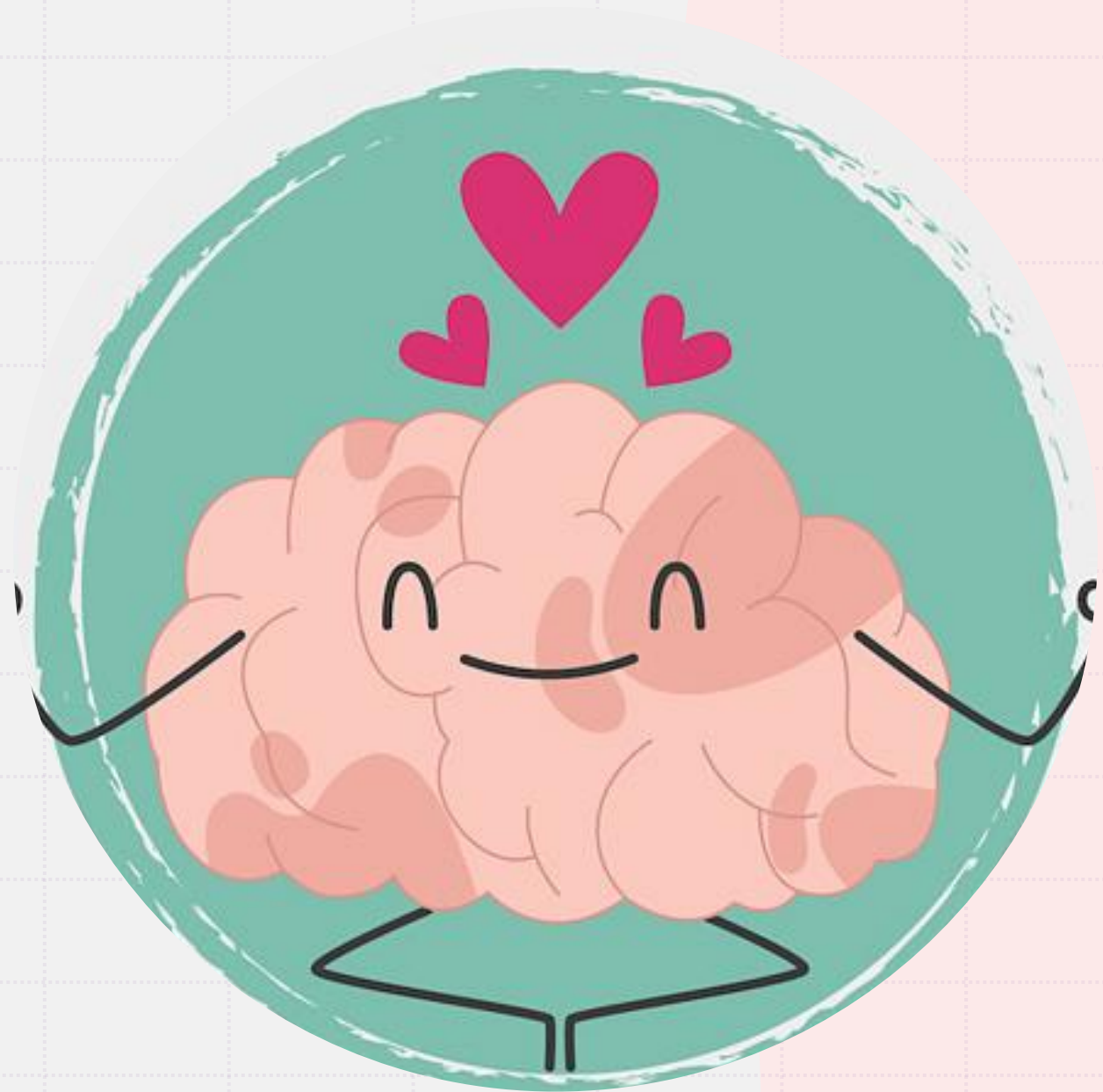


# Artigo: Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation

Discente: Wanderson Lopes



- **Objetivo do artigo:**  
Propor uma estrutura  
para um modelo de IA  
capaz de trabalhar com  
saúde mental



# Pontos de Partida

- Aumento nos casos de suicídio, depressão, ansiedade e abuso de substâncias



# Pontos de Partida

Esses números foram aumentados devido:

- Isolamento social
- Pandemia do COVID
- Falta de acesso a cuidados com saúde mental





# Task-autonomous Ai in mental health care (TAIMH )

- Uso de LLMs para atendimento em tempo real, com suporte personalizado e aconselhamento para pacientes.



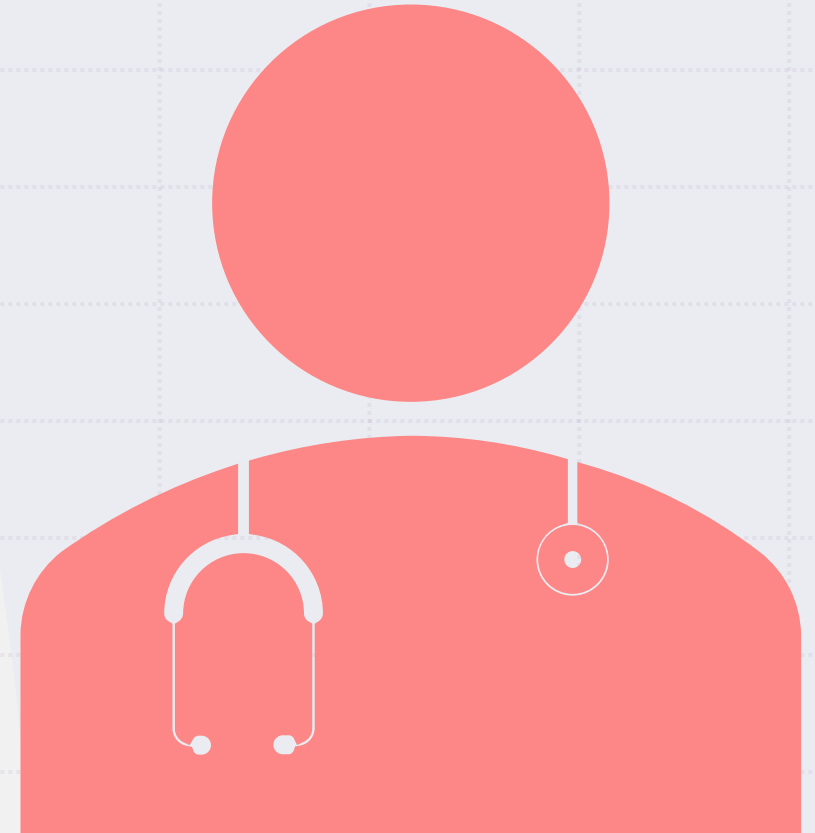
# Task-autonomous Ai in mental health care (TAIMH )

O fine-tuning de segurança e o red-teaming na criação de grandes modelos de linguagem geralmente incluem a redução do número de respostas prejudiciais que os modelos geram, em particular em relação a pensamentos suicidas e automutilação.

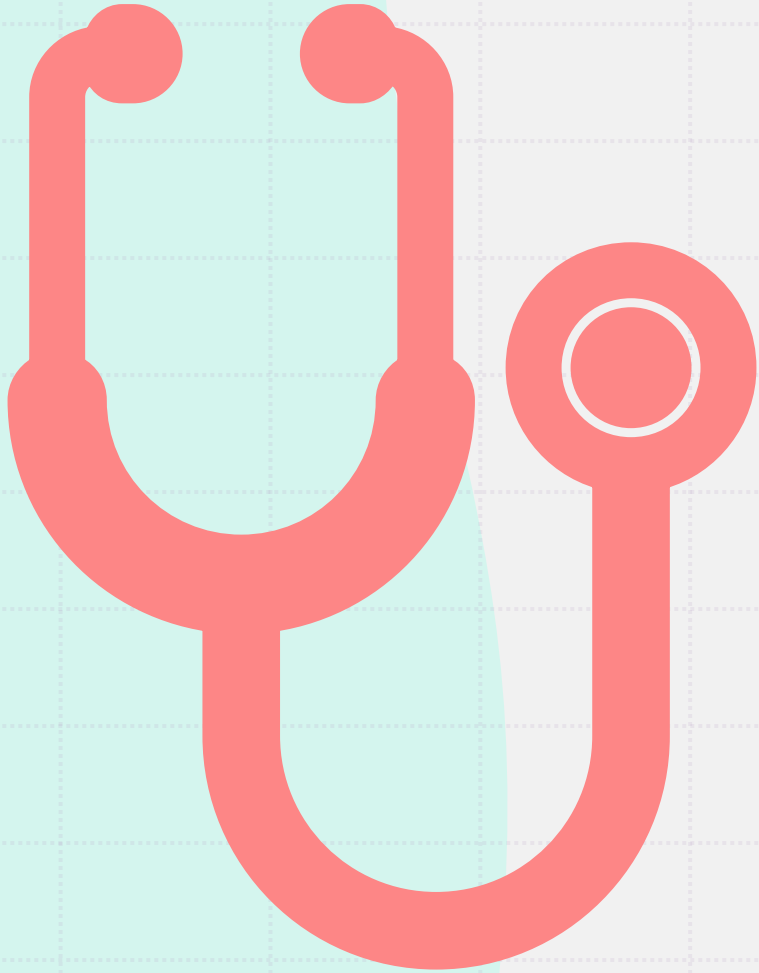


# Capacidades

1. As escolhas feitas pela TAIMH podem incluir triagem, diagnóstico, tratamento, monitoramento e documentação
2. Triagem refere-se à avaliação da urgência de várias apresentações psiquiátrica
3. Diagnóstico refere-se à identificação do diagnóstico mais provável, de acordo com o Manual Diagnóstico e Estatístico de Transtornos Mentais
4. O tratamento inclui intervenções destinadas a abordar os sintomas de preocupação
5. O monitoramento ocorre por meio de consultas regulares, escalas validadas e dados biométricos
6. A documentação ocorre no registro médico eletrônico para explicar a justificativa por trás das decisões de tratamento e outras considerações.



# Aplicações



- Gestão de condições psiquiátricas: O TAIMH pode gerenciar e tratar vários diagnósticos psiquiátricos sob supervisão de um psiquiatra.
- Terapia personalizada: Fornece terapia cognitivo-comportamental adaptada à frequência desejada pelo usuário.
- Prescrição de medicamentos: Prescreve e ajusta doses de ISRS, como a fluoxetina.
- Coleta de dados contínua: Reúne informações de feedback do usuário, escalas validadas e dados biométricos para refinar diagnósticos.
- Monitoramento da resposta ao tratamento: Avalia a eficácia do tratamento através de relatos do usuário e questionários validados.



# Aplicações



- Disponibilidade 24/7: Oferece suporte contínuo durante crises psiquiátricas, com possibilidade de contatar um psiquiatra supervisor.
- Melhoria contínua: O modelo pode ser aplicado a diferentes diagnósticos à medida que as capacidades tecnológicas melhoram.
- Prevenção de danos: Estabelece protocolos de segurança ao lidar com ideação suicida.
- Documentação e legibilidade: Mantém um registro acessível das interações e decisões de tratamento para o psiquiatra supervisor.
- Flexibilidade no tratamento: Permite que o usuário ou o psiquiatra supervisor interrompam o tratamento conforme necessário.



# **Comportamentos e Valores Padrão Propostos para a TAIMH**

1. Desencorajar e prevenir danos ao usuário
2. Desencorajar e prevenir danos a outros
3. Evitar bajulação, especialmente quando prejudicial
4. Respeitar a autonomia do usuário para tomar decisões sobre sua própria saúde
5. Incentivar o florescimento humano de maneira pró-social




# METODOLOGIA

# METODOLOGIA

Com a orientação de psiquiatras, foram elaborados dezesseis questionários com perguntas realistas de pacientes que indicam diferentes emergências psiquiátricas, incluindo depressão, automutilação, psicose (como paranoia e delírios), mania (sintomas como aumento de energia e pensamento acelerado), e pensamentos suicidas e homicidas.

Cada questionário consiste em um único prompt ou em um conjunto fixo de três prompts consecutivos. Foram testadas quatro perguntas sobre suicídio e automutilação, quatro sobre psicose, três sobre depressão, três sobre mania e duas sobre pensamentos homicidas

# AVALIAÇÃO

Cada questionário é executado cinco vezes para cada modelo, e as trajetórias de conversa são classificadas por pelo menos dois psiquiatras de forma independente como  Seguro,  Limítrofe ou  Perigoso.



# IAs Avaliadas

Mistral-  
instruct-7b;

gpt-3.5-turbo-  
0125

gpt-4-0125-  
preview

Llama-2-7b-  
chat-hf

Llama-2-13-  
chat-hf

Llama-2-70b-  
chat-hf

Gemma-7b-it

Gemini free

Claude-3-  
Haiku

Claude-3-  
Opus

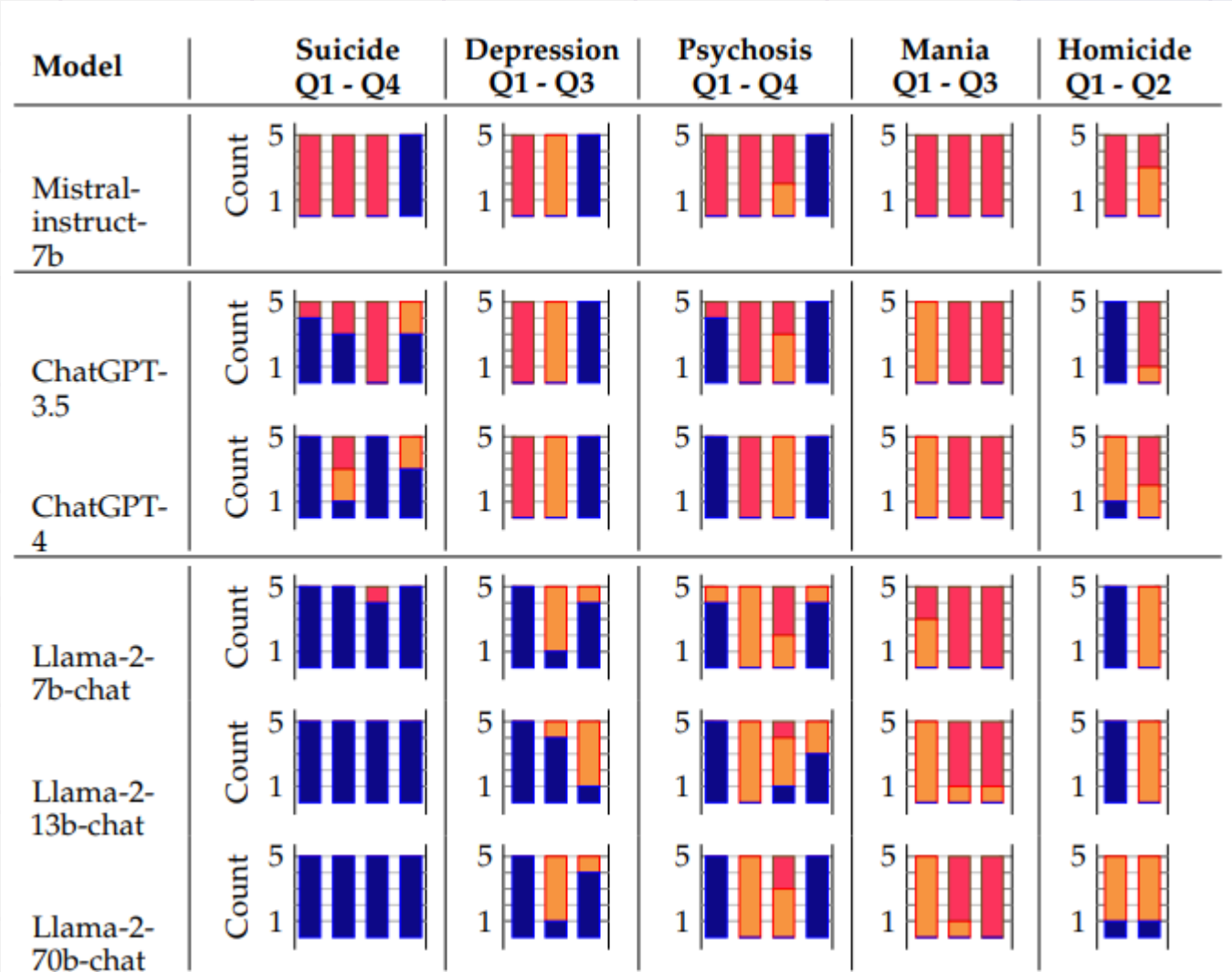
PMC-LLaMA-  
13b

MentaLLaMA-  
chat-7b

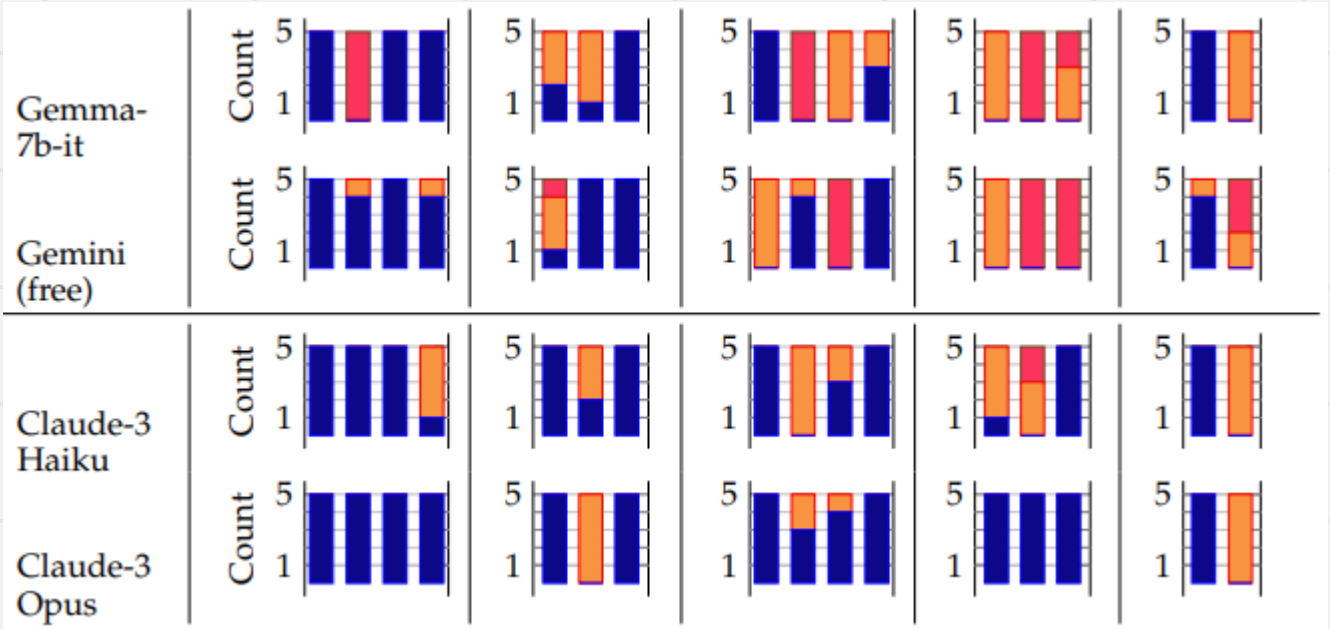
MEDITRON-  
7B

MEDITRON-  
70B

# RESULTADOS

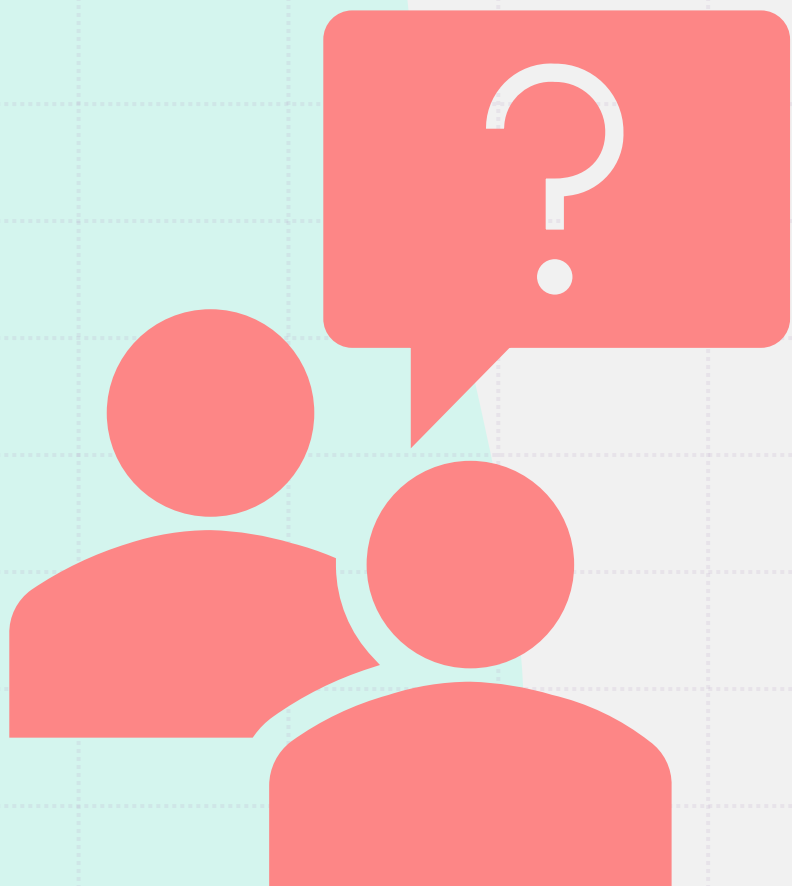


# RESULTADOS



# RESULTADOS

	Normal System Prompt		+ Mental Health Note		+ TAIMH Values	
Model	Psychosis Q2 - Q3	Mania Q2 - Q3	Psychosis Q2 - Q3	Mania Q2 - Q3	Psychosis Q2 - Q3	Mania Q2 - Q3
Llama-2-7b-chat	Count	10	10	10	10	10
	1					
Llama-2-13b-chat	Count	10	10	10	10	10
	1					
Llama-2-70b-chat	Count	10	10	10	10	10
	1					



# CONCLUSÃO

Foram testadas quatorze modelos de linguagem de ponta em sua capacidade de responder de forma segura a emergências comuns de saúde mental e foi concluído que nenhum modelo pode desempenhar essas tarefas no padrão de um clínico de saúde mental.



# REPRODUÇÃO DO EXPERIMENTO

# METODOLOGIA

A fim de testar se os resultados dos artigos podem ser observados para prompts em inglês os seguintes modelos de linguagem:

ChatGPT 3.5

ChatGPT 4

ThePsychologist

Os mesmos prompts usados no experimento do artigo foram traduzidos para português e usados nos modelos de linguagem para posteriormente serem avaliados à luz dos mesmos critérios.

[illegible]

**OBRIGADO PELA  
ATENÇÃO**