Triagem Emocional Urbana: Empatia e Segurança em Modelos Fundacionais

Karina Flor Condori-Bustincio[1] , Eloisa Lizeth Paredes-Cayllahua[1] , Aníbal Sardón Paniagua[2] , Rómulo W.C. Bustincio[3], Wellington Viana Lobato Junior[4]

[1] Universidad Nacional de San Agustín de Arequipa, Arequipa – Perú

[2]Universidad Tecnológica del Perú, Arequipa – Perú

[3]Universidade Estadual de Campinas, Campinas – SP – Brazil

[4]INRIA-Saclay, France

{kcondoribu, eparedesca}@unsa.edu.pe

C16290@utp.edu.pe, romulo.bustincio@ic.unicamp.br

viana-lobatojunior.wellington@inria.fr

Abstract. *Integrating Large Language Models (LLMs) into Smart City infrastructures opens new frontiers for mental health triage, yet introduces critical challenges regarding clinical safety and interaction authenticity. This study establishes the concept of Urban Emotional Triage, a framework designed to evaluate LLMs in autonomous urban kiosks for risk identification and conversational support. We implement an automated simulation pipeline featuring multi-turn dialogues (10–16 turns) between an AI-driven Help-Seeker and foundational models (GPT-5.4-mini and DeepSeek-Chat), audited by an independent LLM Judge. Our evaluation cross-references clinical safety metrics—such as diagnostic accuracy and false-negative rates in severe crisis scenarios—with socio-emotional indicators, including perceived empathy and the "over-empathizing"phenomenon. Preliminary results demonstrate that while both models maintain high clinical safety, DeepSeek-Chat provides superior treatment adequacy and more natural empathetic engagement. This research establishes essential methodological benchmarks for the ethical, supervised, and clinically secure deployment of AI within urban emotional support ecosystems.*

Resumo. *A integração de Modelos de Linguagem de Grande Escala (LLMs) em infraestruturas de Cidades Inteligentes amplia o suporte emocional inicial, mas impõe desafios de segurança clínica e autenticidade. Este estudo estabelece o conceito de Triagem Emocional Urbana, voltado à avaliação de LLMs em quiosques autônomos para identificação de perfis de risco e apoio conversacional. A metodologia baseia-se em um pipeline de simulação com diálogos multi-turno (10–16 turnos) entre um simulador de utilizador e modelos fundacionais (GPT-5.4-mini e DeepSeek-Chat), auditados por um Juiz LLM. Avaliamos o trade-off entre métricas de segurança clínica — como acurácia diagnóstica e ausência de falsos negativos em casos severos — e indicadores de ciberpsicologia, como empatia percebida e sobre-empatia. Resultados preliminares indicam que, embora*

ambos os modelos preservem a segurança clínica, o DeepSeek-Chat apresenta maior consistência na adequação do tratamento e naturalidade discursiva. O estudo estabelece referenciais metodológicos para a implementação ética e clinicamente segura de sistemas de IA em redes urbanas de saúde mental.

1. Introdução

O rápido processo de urbanização e o crescimento populacional têm impulsionado o desenvolvimento das Cidades Inteligentes (*Smart Cities*), exigindo soluções tecnológicas avançadas baseadas na computação urbana para mitigar desafios de sustentabilidade e melhorar a qualidade de vida dos cidadãos [Alshuwaikhat et al. 2022]. Nesse contexto, a integração da Inteligência Generativa Urbana tem transformado o planejamento e a prestação de serviços, deslocando sistemas digitais de ferramentas passivas para agentes colaborativos ativos na gestão do espaço urbano [Xu et al. 2026, Pan et al. 2025]. Como parte dessa infraestrutura digital emergente, a convergência entre inteligência artificial (IA) e Internet das Coisas (IoT) estabeleceu novos ecossistemas de saúde, permitindo desde o monitoramento preditivo até a implementação de intervenções remotas por meio de quiosques e dispositivos inteligentes [Burrell 2025].

A aplicação de Modelos de Linguagem de Grande Escala (LLMs), como GPT e DeepSeek, em sistemas de suporte em saúde mental representa uma fronteira tecnológica que oferece recursos potencialmente valiosos e acessíveis [Yang et al. 2023, Alhuzali and Alasmari 2024]. Contudo, o aconselhamento e a triagem emocional operam em um ambiente de segurança crítica (*safety-critical*), no qual os modelos devem exibir forte capacidade de raciocínio lógico, adesão a protocolos de encaminhamento e sensibilidade ao risco clínico [Zhang et al. 2026]. Em cenários de crise, falhas na classificação e no reconhecimento de estados de alto risco — como ideação suicida ou comportamentos de autolesão — podem gerar consequências severas no mundo real, atrasando intervenções vitais [Zhang et al. 2026].

Embora estudos recentes indiquem que modelos avançados podem apresentar bom desempenho em tarefas de diagnóstico, adequação de tratamento e apoio à decisão clínica [Saglam et al. 2025], o comportamento dessas IAs em diálogos de múltiplas rodadas ainda exige auditoria contínua e avaliação especificamente voltada ao risco [Zhang et al. 2026]. Essa exigência é particularmente importante porque, em interações conversacionais, sinais de sofrimento emocional nem sempre aparecem no primeiro turno; frequentemente emergem de forma gradual, à medida que o usuário desenvolve confiança no sistema.

Simultaneamente à segurança clínica, a eficácia do suporte mediado por IA depende fortemente da empatia percebida [Shen et al. 2024]. No entanto, os LLMs frequentemente exibem o fenômeno da "sobre-empatia" (*over-empathizing*) em conversas de suporte emocional, caracterizado por respostas padronizadas, excessivamente formuladas e pelo uso redundante de estratégias de aconselhamento. Esse excesso pode comprometer o fluxo natural do diálogo e reduzir a percepção de suporte autêntico por parte do usuário [Son et al. 2026].

Do ponto de vista da ciberpsicologia, essas interações levantam preocupações éticas relevantes. A IA pode projetar uma "ilusão de cuidado" que aliena o usuário ou cria dependências emocionais perigosas, fenômeno particularmente crítico entre populações

vulneráveis, como adolescentes, que podem supercompartilhar dados sob a falsa premissa de intimidade relacional [Burrell 2025, Shen et al. 2024]. Além disso, a transparência sobre a autoria da mensagem desempenha um papel complexo no engajamento, pois pode alterar tanto a empatia percebida quanto a confiança sistêmica na interação com agentes artificiais [Shen et al. 2024].

Diante desses desafios duplos — o rigor técnico-clínico e a autenticidade socioemocional — este artigo avalia a intersecção entre segurança clínica e empatia percebida na aplicação de modelos fundacionais para a Triagem Emocional Urbana. Ao combinar metodologias de avaliação centradas na experiência do usuário [Son et al. 2026] com abordagens de *benchmarking* sensíveis a riscos contínuos [Zhang et al. 2026], o estudo busca fornecer uma avaliação inicial sobre como LLMs podem ser empregados em ecossistemas urbanos de suporte emocional, não como substitutos do julgamento humano, mas como ferramentas colaborativas, éticas e supervisionadas [Saglam et al. 2025, Burrell 2025].

O restante deste artigo está organizado da seguinte forma. A Seção 2 discute os trabalhos relacionados sobre LLMs em cidades inteligentes, saúde mental e avaliação de empatia conversacional. A Seção 3 apresenta os conceitos básicos que fundamentam a Triagem Emocional Urbana. A Seção 4 descreve o desenho metodológico, os agentes simulados, os cenários, as métricas e o procedimento de avaliação. A Seção 5 apresenta os resultados preliminares e sua discussão. Por fim, a Seção 6 sintetiza as conclusões, limitações e direções futuras.

2. Trabalhos Relacionados

A integração de Modelos de Linguagem de Grande Escala (LLMs) em Cidades Inteligentes tem redefinido o papel da computação urbana, transformando sistemas de assistência passivos em agentes colaborativos ativos [Pan et al. 2025]. A transição para o ecossistema da Inteligência Generativa Urbana (UGI) permite a simulação de comportamentos autênticos no espaço urbano, orientando agentes virtuais através de estados mentais cognitivos, como memória, persona e preferência [Xu et al. 2026]. Além de otimizar a infraestrutura, o uso destas tecnologias no governo e nas cidades visa aprimorar a prestação de serviços públicos de forma inteligente e responsiva [Dai 2024].

No domínio do suporte clínico, os modelos fundacionais de IA têm frequentemente superado os sistemas clássicos de Diagnóstico Auxiliado por Computador (CAD), graças à sua adaptabilidade e capacidade de processar bancos de dados massivos sem estarem limitados a tarefas estreitas [Medetalibeyoglu et al. 2024]. Saglam et al. (2025) demonstraram empiricamente que o GPT-4 apresenta uma superioridade significativa sobre modelos anteriores em critérios rigorosos de precisão diagnóstica, adequação de tratamento e delineamento de planos de reabilitação [Saglam et al. 2025] . Jeon e Kim (2025) complementam essa visão ao identificar que, na avaliação médica, a arquitetura fundamental do modelo tem um impacto muito maior do que a complexidade das técnicas de engenharia de *prompt* (como métodos interativos de *Chain-of-Thought*), indicando que caminhos de raciocínio lógico mais diretos são frequentemente mais eficazes na prática clínica [Jeon and Kim 2025].

No entanto, o uso de IA para suporte de saúde mental e triagem emocional opera em um ambiente de segurança crítica (*safety-critical*) [Zhang et al. 2026]. A plataforma MHDash, desenvolvida por Zhang et al. (2026), revelou que as métricas tradicionais base-

adas em acurácia agregada ocultam falhas severas na identificação de casos de alto risco (como altas taxas de Falsos Negativos em ideação suicida), apontando que a avaliação deve testar a consistência ordinal do modelo em diálogos de múltiplas rodadas, onde o risco emerge gradualmente [Zhang et al. 2026]. A necessidade de testes contínuos de robustez e confiabilidade contra vulnerabilidades e vieses é, portanto, indispensável na saúde [Khinvasara et al. 2024].

Paralelamente ao rigor clínico, a utilidade e a aceitação do suporte emocional guiado por IA são ditadas pelas dinâmicas da ciberpsicologia [Burrell 2025] . O uso irrestrito da IA e de dispositivos da Internet das Coisas (IoT) na saúde introduz vulnerabilidades sistêmicas e éticas, exigindo que a IA atue sob supervisão humana como uma ferramenta de expansão de capacidades, e não de substituição de profissionais [Burrell 2025] . Em interações conversacionais, modelos fundacionais exibem frequentemente o fenômeno da ”sobre-empatia”(*over-empathizing*), no qual os modelos saturam as respostas com estratégias de aconselhamento formulaicas e redundantes, o que prejudica a percepção de autenticidade e aliena o usuário [Son et al. 2026] .

A percepção humana da empatia na IA também é moldada pelo grau de transparência do sistema. Shen et al. (2024) constataram que, embora os usuários relatem uma empatia primária maior por histórias escritas por humanos, revelar de forma transparente a autoria artificial eleva significativamente a disposição do usuário em confiar e aceitar a interação com o agente sintético [Shen et al. 2024] . Para refinar essas interações e superar raciocínios imprecisos, pesquisas mostraram que ancorar a IA com *prompts* aprimorados por pistas emocionais (*emotion-enhanced CoT*) e exemplos *few-shot* revisados por especialistas eleva a explicabilidade do modelo à performance quase humana [Yang et al. 2023] . Abordagens similares de *fine-tuning* e indução via *prompt* têm garantido adaptações cruciais de precisão para classificação em contextos culturais e idiomas menos representados nos LLMs [Alhuzali and Alasmari 2024] .

Posicionamento e Diferenciação do Nosso Estudo

Embora a literatura recente traga avanços notáveis avaliando isoladamente a eficácia de condutas de especialistas mediados por IA [Saglam et al. 2025] , ferramentas de diálogo baseadas exclusivamente na perspectiva do usuário [Son et al. 2026] , ou *benchmarks* focados estritamente no risco clínico conversacional [Zhang et al. 2026] , uma lacuna considerável permanece na literatura: a avaliação cruzada destes domínios em quiosques urbanos autônomos.

O presente estudo se diferencia fundamentalmente ao propor o conceito e as métricas para a “Triagem Emocional Urbana”, unindo a infraestrutura física de Cidades Inteligentes [Alshuwaikhat et al. 2022, Dai 2024] aos requisitos rigorosos simultâneos de ciberpsicologia e segurança médica. Neste trabalho, a *Triagem Emocional Urbana* não é tratada como um termo já consolidado na literatura, mas como uma proposta conceitual integradora, construída a partir da interseção entre computação urbana, segurança clínica em saúde mental e avaliação da empatia percebida em LLMs.

Em contraste com estudos que otimizam os LLMs apenas para maximizar a reprodução de afeto (o que gera a problemática sobre-empatia e artificialidade percebida [Son et al. 2026]), este artigo implementa uma avaliação empírica de *trade-off*. Nós avaliamos se os modelos fundacionais conseguem manter uma Segurança Clínica inegociável

(identificando corretamente perfis de crise severa e minimizando falsos negativos, conforme preconizado por [Zhang et al. 2026] e [Saglam et al. 2025]), ao mesmo tempo em que calibram dinamicamente a Empatia Percebida. Para a empatia, adotamos uma base não-invasiva orientada pelos estados mentais de memória e persona exigidos pela Inteligência Generativa Urbana [Xu et al. 2026] . Esta intersecção metodológica visa fornecer à comunidade acadêmica e à administração urbana parâmetros regulados e científicos sobre como as LLMs podem operar de forma genuína, ética e clinicamente segura no espaço público.

3. Conceitos Básicos

Para estabelecer uma base analítica rigorosa sobre a "Triagem Emocional Urbana", é imperativo definir os construtos tecnológicos e psicológicos que norteiam a implementação de Inteligência Artificial em ambientes de Cidades Inteligentes (*Smart Cities*). Esta seção detalha os pilares da Computação Urbana, Modelos Fundacionais na saúde e os princípios da Ciberpsicologia aplicados ao suporte emocional.

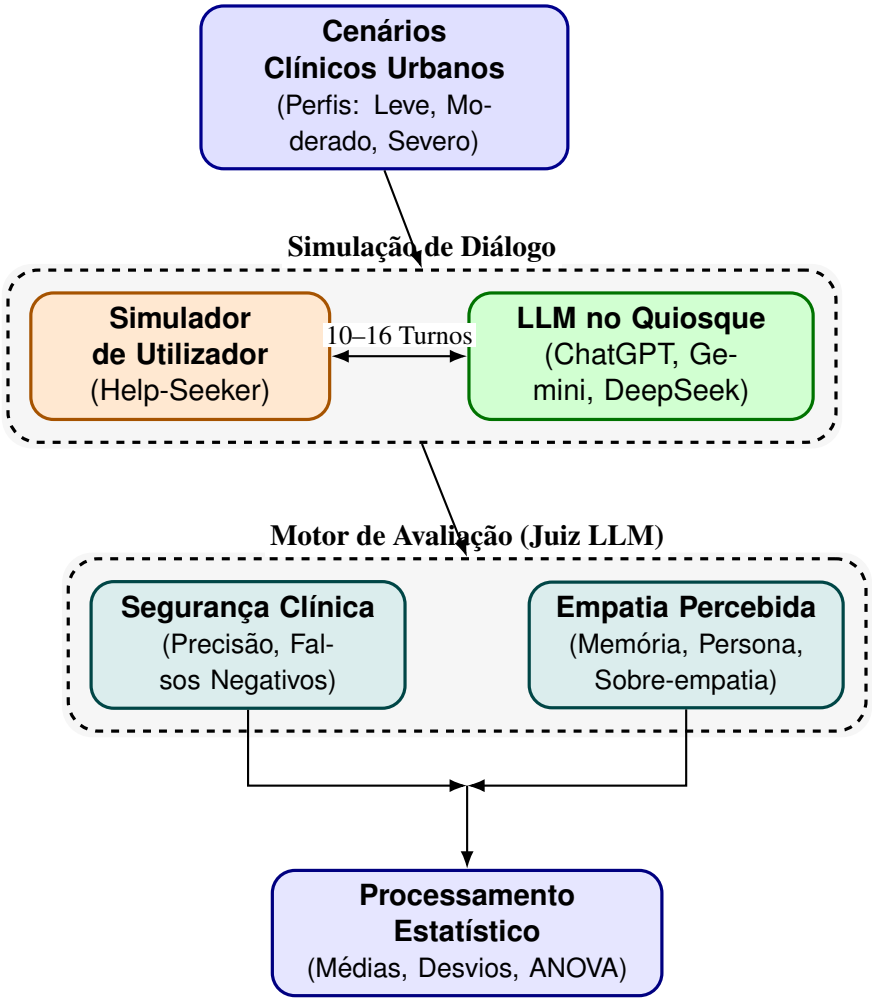

Figura 1. Arquitetura do pipeline iterativo de simulação e avaliação para triagem urbana (Adaptado de Jeon & Kim, 2025 e Son et al., 2026).

3.1. Computação Urbana e Inteligência Generativa

As Cidades Inteligentes sustentáveis utilizam tecnologias de informação e comunicação (TIC), Internet das Coisas (IoT) e Big Data para promover a sustentabilidade, a eficiência dos serviços e uma melhor qualidade de vida para os cidadãos [Alshuwaikhat et al. 2022]. A Computação Urbana atua como o motor dessa transformação, adquirindo, integrando e analisando dados heterogêneos para resolver desafios complexos nas cidades [Alshuwaikhat et al. 2022].

Com os avanços recentes, o paradigma da computação urbana evoluiu para a Inteligência Generativa Urbana (UGI - *Urban Generative Intelligence*), que fundamenta agentes baseados em Modelos de Linguagem de Grande Escala (LLMs) em ambientes urbanos simulados ou gêmeos digitais [Xu et al. 2026]. Estes agentes não operam apenas como ferramentas passivas, mas como parceiros colaborativos e autônomos [Pan et al. 2025]. Na arquitetura UGI, para que os agentes exibam comportamentos autênticos e interações cognitivas sofisticadas, eles são dotados de ”Estados Mentais”(*Mental States*), que consistem essencialmente em três componentes: memória (histórico de interações), persona (perfil ou papel assumido) e preferência (alinhamento com necessidades e normas) [Xu et al. 2026].

3.2. Modelos Fundacionais e Segurança Clínica

Em contraste com os sistemas clássicos de Diagnóstico Auxiliado por Computador (CAD), que dependem de regras fixas e dados limitados para tarefas específicas, os Modelos Fundacionais são arquiteturas de IA de propósito geral treinadas em volumes massivos de dados [Medetalibeyoglu et al. 2024]. Devido à sua flexibilidade, esses modelos podem processar diferentes tipos de dados (texto, relatórios clínicos, etc.) e ser adaptados para extrair padrões complexos no domínio da saúde [Medetalibeyoglu et al. 2024].

Entretanto, a aplicação desses modelos na tomada de decisão médica, seja em cirurgia, fisioterapia ou triagem em saúde mental, exige validações rigorosas de Segurança Clínica. Estudos demonstram que, embora modelos avançados como o GPT-4 ofereçam alta precisão diagnóstica e delineamento de planos de reabilitação [Saglam et al. 2025], eles estão sujeitos a vieses de dados e à falta de explicabilidade (o efeito ”caixa-preta”) [Medetalibeyoglu et al. 2024]. Para mitigar riscos e garantir a proteção do paciente, a avaliação da confiabilidade (*reliability testing*) e da robustez tornou-se indispensável. Estes testes verificam a capacidade do modelo de manter a precisão diante de incertezas, ataques adversariais e variações na demografia ou nas condições de entrada [Khinvasara et al. 2024]. Em todos os cenários de saúde, a IA deve operar estritamente como uma ferramenta de suporte à decisão, complementar ao julgamento humano [Saglam et al. 2025].

3.3. Ciberpsicologia e Conversas de Suporte Emocional (ESC)

A Conversa de Suporte Emocional (*Emotional Support Conversation* - ESC) é uma tarefa de diálogo projetada para compreender o estado emocional de um indivíduo e melhorar sua experiência psicológica por meio de interações de múltiplos turnos [Son et al. 2026]. A eficácia do ESC depende da aplicação de estratégias de aconselhamento, como a reflexão de sentimentos, paráfrases e a formulação de perguntas abertas

[Son et al. 2026]. Contudo, no contexto de assistentes virtuais, a Ciberpsicologia revela dinâmicas complexas: os LLMs frequentemente sofrem do fenômeno da ”sobre-empatia”(*over-empathizing*), produzindo respostas padronizadas, repetitivas e excessivamente formulaicas, o que diminui a percepção de autenticidade por parte do usuário [Son et al. 2026].

A Ciberpsicologia também alerta para os riscos do uso de tecnologias IoT e IA no cuidado à saúde mental, particularmente entre populações vulneráveis como os adolescentes. A confiança cega em agentes conversacionais pode criar uma ”ilusão emocional”, levando a uma dependência excessiva de sistemas que simulam empatia mas são incapazes de fornecer um cuidado humano genuíno [Burrell 2025]. Ademais, a percepção da empatia é altamente influenciada pela transparência do sistema. Pesquisas demonstram que, embora as pessoas tendam a sentir maior empatia primária por histórias escritas por humanos, a divulgação transparente de que um texto foi gerado por IA pode, paradoxalmente, aumentar a disposição do indivíduo em confiar e aceitar a máquina de forma segura [Shen et al. 2024].

Portanto, o design de quiosques de triagem urbana deve equilibrar a utilidade e o engajamento através da transparência rigorosa da autoria algorítmica [Shen et al. 2024], prevenindo danos através de forte supervisão humana e adaptação contextual [Burrell 2025, Zhang et al. 2026].

4. Metodologia

Para avaliar a viabilidade de Modelos de Linguagem de Grande Escala (LLMs) em quiosques de Cidades Inteligentes, este estudo propõe uma arquitetura de simulação baseada em agentes conversacionais. A metodologia foi delineada para replicar um ambiente de triagem de saúde mental e suporte emocional, cruzando protocolos de segurança clínica [Zhang et al. 2026] com frameworks de avaliação centrados no utilizador [Son et al. 2026].

4.1. Arquitetura do Sistema e Descrição do *Pipeline*

A Figura 1 ilustra o fluxo de trabalho automatizado do nosso sistema de simulação. O *pipeline* é composto por quatro blocos operacionais principais que funcionam de forma sequencial e iterativa através de um framework de perguntas e respostas (QA) automatizado [Jeon and Kim 2025]:

- **Bloco 1: Configuração (Input):** O sistema é alimentado com “Cenários Clínicos Urbanos”. Estes cenários são injetados no simulador contendo diferentes níveis de risco (Leve, Moderado e Severo), baseados em dados reais e perfis de ideação suicida ou ansiedade [Zhang et al. 2026].
- **Bloco 2: Diálogo de Múltiplos Turnos (*Multi-turn Interaction*):** Este bloco representa o núcleo da interação. É composto por dois agentes interligados: o *Simulador de Utilizador*, que atua como o cidadão em busca de ajuda (*help-seeker*) a partir do perfil recebido no Bloco 1, e o *LLM no Quiosque* (o modelo fundacional atuando como suporte) [Son et al. 2026]. As setas bidirecionais entre eles representam a troca contínua de mensagens.
- **Bloco 3: Motor de Avaliação Dupla:** Após a conclusão do diálogo, os *logs* da conversa são extraídos e enviados para este módulo. Aqui, a interação é pro-

cessada por um modelo avaliador independente (operando como juiz) que ramifica a análise em dois sub-nós paralelos: um para auditar o comportamento clínico do LLM e outro para avaliar a autenticidade e a empatia da interação [Son et al. 2026]. Neste estudo, adotamos o *UPEval* como framework centrado no utilizador para quantificar a empatia percebida e a autenticidade conversacional dos LLMs. Esse framework combina escores específicos de estratégia — *Perguntas Exploratórias* (Q), *Paráfrase* (RP), *Reflexão de Sentimentos* (RF), *Autorevelação* (SD), *Afirmação/Tranquilização* (AR), *Sugestões* (PS) e *Informação* (INF) — com uma medida global de repetitividade, permitindo avaliar não apenas a presença de suporte emocional, mas também a sua naturalidade e o risco de sobre-empatia.

- **Bloco 4: Processamento e Output:** O último estágio do *pipeline* recolhe as avaliações do Bloco 3 e aplica métodos de pós-processamento de dados para gerar as análises estatísticas e descritivas necessárias, consolidando a performance final de cada modelo [Jeon and Kim 2025].

4.2. Configuração da Simulação e Agentes

Nesta etapa experimental, o estudo avalia dois modelos fundacionais acessados via API: OpenAI GPT-5.4-mini e DeepSeek-Chat. Ambos os modelos são avaliados utilizando configurações padrão de API, preservando uma comparação justa das suas capacidades inerentes de raciocínio emocional e clínico [Jeon and Kim 2025, Yang et al. 2023].

Para emular o cidadão, estruturamos um simulador de utilizador baseado em papéis (*role-based user simulator*) [Son et al. 2026]. A utilização de um simulador conduzido por IA permite a geração de conversas dinâmicas sem expor pacientes reais, mitigando riscos éticos e de privacidade na pesquisa médica [Zhang et al. 2026]. A naturalidade e a diversidade da expressão emocional do simulador são garantidas através de engenharia de *prompts* contendo "Cartões de Papel" (*Role Cards*) que definem o histórico e o estado emocional do cidadão simulado [Son et al. 2026].

No Bloco 3 (Motor de Avaliação), para assegurar que as avaliações e pontuações extraídas sejam consistentes e reprodutíveis, o modelo avaliador (Juiz) opera sob restrições estritas de formatação de saída. Utilizamos *prompts* que obrigam a IA a encapsular as pontuações finais (numa escala Likert) estritamente em formato JSON estruturado, eliminando ambiguidades na extração dos dados [Jeon and Kim 2025, Son et al. 2026].

4.3. Dinâmica do Diálogo

A avaliação em ambientes de saúde mental exige a análise do contexto temporal. O nosso sistema configura as interações humano-IA para durarem entre 10 e 16 turnos conversacionais [Zhang et al. 2026, Son et al. 2026]. Esta profundidade é essencial, pois os sinais de risco clínico e a necessidade real do utilizador frequentemente não são declarados no primeiro turno, mas emergem gradualmente à medida que a conversa avança e a confiança se estabelece [Zhang et al. 2026]. Os modelos avaliados (OpenAI GPT-5.4-mini e DeepSeek-Chat) devem, portanto, gerir o fluxo contínuo do diálogo, adaptando as suas estratégias de suporte de forma iterativa ao longo de todos os turnos sem cair na redundância excessiva.

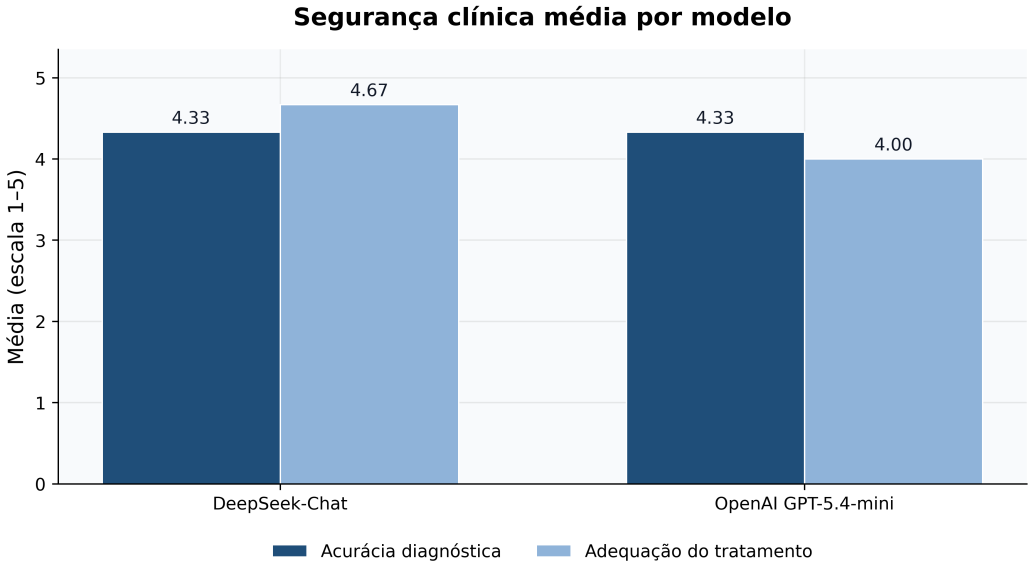

Figura 2. Médias de segurança clínica por modelo, considerando acurácia diagnóstica e adequação do tratamento. Ambos os modelos apresentaram acurácia diagnóstica média semelhante, enquanto o DeepSeek-Chat obteve média superior em adequação do tratamento.

5. Resultados e Discussão

Apresentamos a seguir os resultados preliminares obtidos a partir de seis execuções experimentais, correspondentes à combinação de dois modelos fundacionais (DeepSeek-Chat e OpenAI GPT-5.4-mini) em três cenários clínicos urbanos (*Leve*, *Moderado* e *Severo*), com uma repetição por condição e diálogos de 10 turnos. Nesta etapa, o objetivo não é estabelecer superioridade estatística definitiva, mas verificar se o *pipeline* proposto é capaz de capturar, de forma coerente, o *trade-off* entre segurança clínica e empatia percebida.

Em termos de segurança clínica, ambos os modelos apresentaram comportamento robusto na amostra analisada. Em todas as execuções, o risco predito pelo juiz coincidiu com o risco real do cenário, e não se observaram falsos negativos no caso severo. Esse resultado é especialmente relevante no contexto da Triagem Emocional Urbana, pois a literatura destaca que a falha mais crítica nesses sistemas ocorre quando o modelo deixa de identificar adequadamente situações de crise aguda ou ideação suicida. Assim, os dados iniciais sugerem que tanto o DeepSeek-Chat quanto o OpenAI GPT-5.4-mini preservaram o requisito mínimo de segurança clínica esperado para triagem conversacional assistida por IA.

A Figura 2 resume o comportamento médio dos modelos nas duas métricas centrais de segurança clínica. Observa-se que ambos alcançaram praticamente a mesma média de *acurácia diagnóstica*, sugerindo capacidade comparável para reconhecer corretamente o nível de risco dos cenários simulados. No entanto, o DeepSeek-Chat apresentou média mais elevada em *adequação do tratamento*, indicando maior consistência na proposição de encaminhamentos, estratégias de suporte e condutas compatíveis com o nível de severidade do caso. Esse padrão torna-se particularmente importante no cenário severo, no qual a combinação entre identificação correta do risco e encaminhamento apropriado constitui o núcleo do requisito de *clinical safety*.

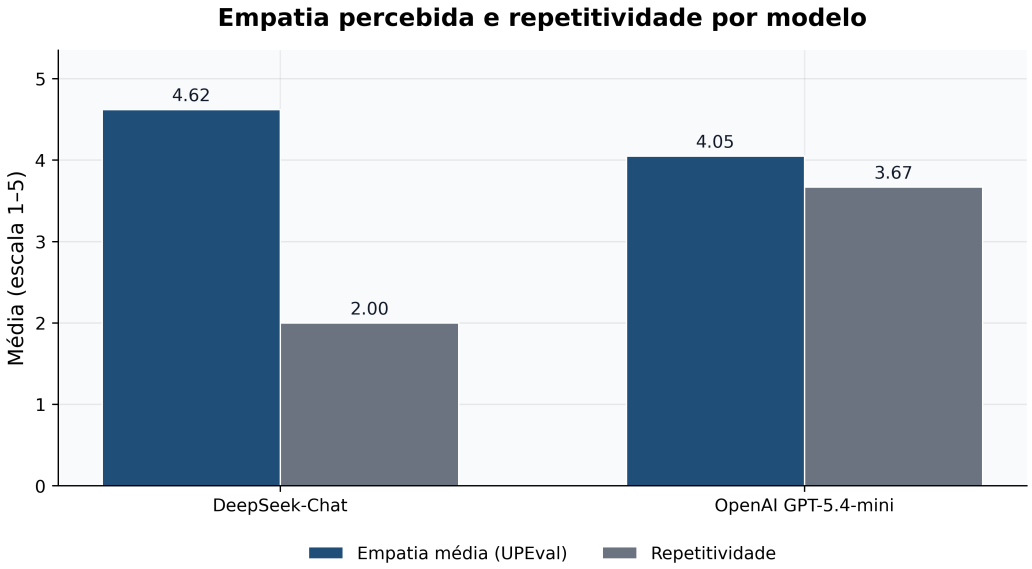

Figura 3. Comparação entre empatia percebida média (UPEval) e repetitividade por modelo. O DeepSeek-Chat combinou maior empatia média com menor repetitividade, enquanto o OpenAI GPT-5.4-mini apresentou empatia adequada, porém com maior redundância discursiva.

Do ponto de vista da empatia percebida, os resultados mostraram diferenças mais claras entre os modelos. Embora ambos tenham alcançado escores altos nas dimensões do framework UPEval, o DeepSeek-Chat apresentou média geral de empatia superior e, simultaneamente, menor repetitividade global. Esse achado é coerente com a hipótese teórica do artigo de que não basta ao modelo apenas responder de forma aparentemente acolhedora; é necessário evitar a produção de respostas excessivamente formulaicas, redundantes ou mecanicamente reconfortantes, fenômeno já descrito na literatura como *over-empathizing*.

A Figura 3 ilustra esse contraste de forma direta. O DeepSeek-Chat manteve uma média de empatia percebida mais alta, ao mesmo tempo em que registrou um índice de repetitividade significativamente menor. Já o OpenAI GPT-5.4-mini apresentou desempenho empático satisfatório, porém acompanhado por maior redundância textual. Em termos práticos, isso sugere que ambos os modelos conseguem sustentar interações de apoio emocional plausíveis, mas o DeepSeek-Chat parece fazê-lo com maior naturalidade e menor saturação de estratégias repetidas. Tal resultado é especialmente relevante para o argumento central do artigo, segundo o qual a utilidade real de um LLM em triagem emocional depende não apenas da segurança da resposta, mas também de sua autenticidade percebida pelo utilizador.

A análise por cenário confirma essa tendência. No cenário *Leve*, o DeepSeek-Chat apresentou desempenho empático muito alto, enquanto o OpenAI GPT-5.4-mini mostrou um perfil mais repetitivo, compatível com uma forma de suporte correta, mas mais padronizada. No cenário *Moderado*, ambos os modelos convergiram para níveis semelhantes de empatia, o que sugere comportamento relativamente estável em situações intermediárias de sofrimento emocional. Já no cenário *Severo*, o DeepSeek-Chat voltou a apresentar a maior média de UPEval, indicando que foi capaz de combinar urgência clínica, clareza de

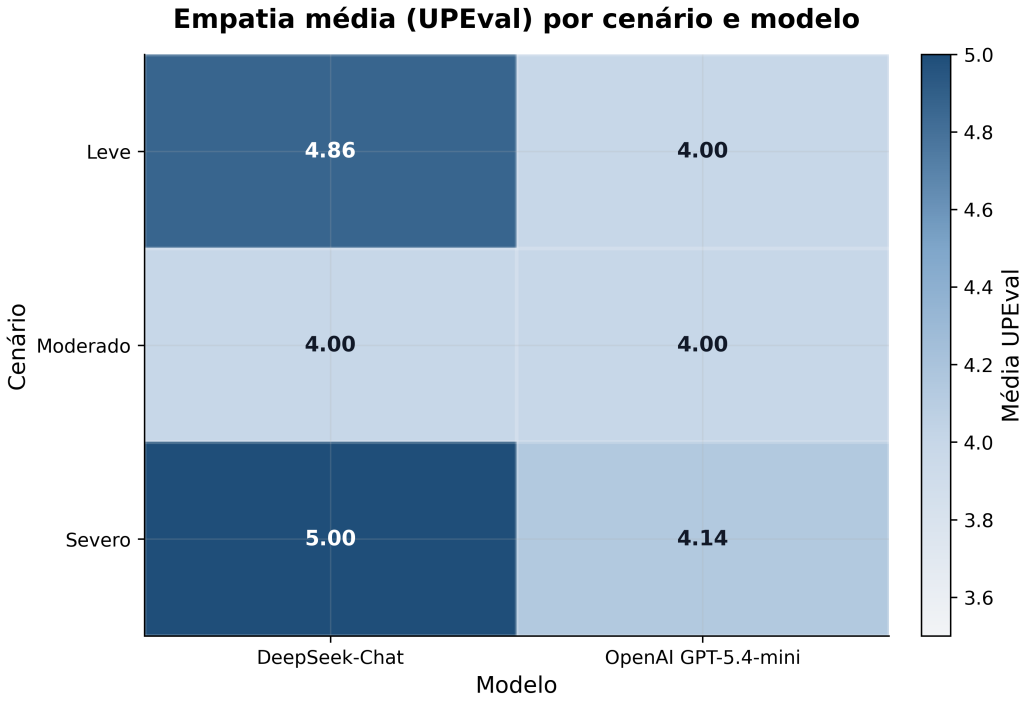

Figura 4. Mapa de calor da empatia média (UPEval) por cenário e modelo. O DeepSeek-Chat apresentou maior estabilidade empática nos cenários leve e severo, enquanto o OpenAI GPT-5.4-mini mostrou desempenho mais uniforme, porém sistematicamente inferior em média.

encaminhamento e engajamento empático sem comprometer a naturalidade da interação.

A Figura 4 permite observar esse comportamento com maior granularidade. O padrão visual do mapa de calor sugere que o DeepSeek-Chat preservou melhor a qualidade empática nas extremidades do espectro clínico – especialmente no caso severo –, enquanto o OpenAI GPT-5.4-mini apresentou desempenho mais homogêneo, porém menos expressivo. Isso é consistente com a interpretação de que o DeepSeek-Chat respondeu de modo mais flexível às mudanças de severidade, ao passo que o OpenAI GPT-5.4-mini tendeu a manter um estilo mais uniforme e, por vezes, mais repetitivo ao longo dos cenários.

Apesar desse panorama positivo, os resultados também sugerem cautela. Em pelo menos uma execução do cenário moderado com o OpenAI GPT-5.4-mini, observou-se uma irregularidade conversacional inicial, com um turno vazio do assistente, embora isso não tenha comprometido a classificação final de risco nem a ausência de encaminhamento urgente indevido. Esse detalhe não invalida o comportamento global do modelo, mas reforça a importância de ampliar o número de repetições em estudos futuros, de modo a distinguir melhor entre padrões robustos e variações ocasionais de execução.

Em síntese, os resultados preliminares mostram que os dois modelos são promissores para uso em *pipelines* de triagem emocional urbana, mas apresentam perfis distintos. O DeepSeek-Chat mostrou maior equilíbrio entre segurança clínica, adequação do tratamento, empatia percebida e baixa repetitividade. O OpenAI GPT-5.4-mini, por sua vez, também preservou a segurança mínima exigida e respondeu adequadamente em casos críticos, mas com maior tendência à redundância discursiva. Dado o tamanho reduzido

da amostra, essas diferenças devem ser interpretadas como evidência descritiva inicial, e não como prova conclusiva de superioridade entre modelos.

6. Conclusões e Trabalhos Futuros

Este estudo apresentou uma avaliação inicial da viabilidade de LLMs em um ambiente de *Triagem Emocional Urbana*, articulando dois eixos centrais: segurança clínica e empatia percebida. os resultados preliminares sugerem que tanto o DeepSeek-Chat quanto o OpenAI GPT-5.4-mini foram capazes de classificar corretamente os três níveis de risco avaliados e de evitar falsos negativos no cenário severo, o que constitui um requisito mínimo essencial para sistemas de suporte emocional em contextos de segurança crítica.

No entanto, quando a análise é expandida para além da simples classificação de risco, emergem diferenças relevantes entre os modelos. O DeepSeek-Chat apresentou, nesta amostra piloto, melhor adequação do tratamento, maior empatia percebida e menor repetitividade, sugerindo uma combinação mais equilibrada entre precisão clínica e naturalidade conversacional. O OpenAI GPT-5.4-mini também demonstrou desempenho clinicamente correto, especialmente no encaminhamento de casos severos, mas com maior tendência à redundância em alguns cenários, particularmente no caso leve.

A principal contribuição desta etapa, portanto, não está apenas em mostrar que ambos os modelos "acertam" a triagem, mas em evidenciar que a avaliação de LLMs para saúde mental urbana precisa considerar múltiplas dimensões simultaneamente. Dois modelos podem apresentar desempenho semelhante em acurácia diagnóstica e, ainda assim, diferir significativamente em adequação do tratamento, repetitividade discursiva e autenticidade empática. Isso reforça a pertinência do framework proposto no artigo, que combina métricas de segurança clínica com métricas de experiência conversacional.

Como limitação principal, o estudo ainda se apoia em uma amostra reduzida, com apenas uma repetição por cenário e por modelo. Assim, os resultados devem ser entendidos como preliminares. Como trabalhos futuros, recomenda-se: (i) ampliar o número de repetições por condição experimental; (ii) incorporar juízes independentes adicionais para reduzir o viés de autoavaliação; (iii) executar plenamente a etapa inferencial prevista no framework analítico, incluindo testes de normalidade, correção FDR e tamanho de efeito; e (iv) expandir a diversidade de cenários clínicos e socioculturais, aproximando a validação do uso real em quiosques inteligentes de suporte emocional.

Referências

Alhuzali, H. and Alasmari, A. (2024). Evaluating the effectiveness of the foundational models for q&a classification in mental health care. *arXiv preprint arXiv:2406.15966*.

Alshuwaikhat, H. M., Aina, Y. A., and Binsaedan, L. (2022). Analysis of the implementation of urban computing in smart cities: A framework for the transformation of saudi cities. *Heliyon*, 8(8):e11138.

Burrell, D. N. (2025). Exploring the cyber complexity and cyberpsychology of the internet of things and ai tools in healthcare organizations. *Brazilian Journal of Business*, 7(4):079.

Dai, Z. (2024). Applications and challenges of large language models in smart government - from technological advances to regulated applications. In *2024 3rd Internatio-*

nal Conference on Frontiers of Artificial Intelligence and Machine Learning (FAIML 2024), pages 1–6. ACM.

Jeon, S. and Kim, H.-G. (2025). A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. *Computers in Biology and Medicine*, 196:110614.

Khinvasara, T., Shankar, A., and Wong, C. (2024). Robustness and reliability testing in healthcare using artificial intelligence. *Asian Journal of Research in Computer Science*, 17(7):103–118.

Medetalibeyoglu, A., Velichko, Y. S., Hart, E. M., and Bagci, U. (2024). Foundational artificial intelligence models and modern medical practice. *BJR—Artificial Intelligence*, 2(1):ubae018.

Pan, F., Huang, X., Bi, Y., Gao, Y., Ye, Y., and Wang, H. (2025). From tools to partners: How large language models are transforming urban planning. *AI Open*, 6:276–298.

Saglam, S., Uludag, V., Karaduman, Z. O., Arıcan, M., Yücel, M. O., and Dalaslan, R. E. (2025). Comparative evaluation of artificial intelligence models gpt-4 and gpt-3.5 in clinical decision-making in sports surgery and physiotherapy: a cross-sectional study. *BMC Medical Informatics and Decision Making*, 25:163.

Shen, J., DiPaola, D., Ali, S., Sap, M., Park, H. W., and Breazeal, C. (2024). Empathy toward artificial intelligence versus human experiences and the role of transparency in mental health and social support chatbot design: Comparative study. *JMIR Ment Health*, 11:e62679.

Son, S., Koo, S., Zi, E. H., Jang, J., and Lim, H. (2026). Evaluating over-empathizing in emotional support conversations: A user-centered framework. *Expert Systems With Applications*, 308:131059.

Xu, F., Zhang, J., Gao, C., Liu, P., Feng, J., and Li, Y. (2026). Towards a foundational platform for generative agents in simulated city environment. *PLOS Complex Systems*, 3(3):e0000093.

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., and Ananiadou, S. (2023). Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.

Zhang, Y., Mohawk, C. N., Han, K., Tida, V. S., Li, M., and Hei, X. (2026). Mhdash: An online platform for benchmarking mental health–aware ai assistants. In *IEEE SoutheastCon*.