Entre Autonomia e Risco: Análise Sociotécnica do Uso de LLMs por Pacientes na Interpretação de Laudos Psicológicos

Dárlinton Barbosa Feres Carvalho, Evaldo de Paula Souza, Bárbara Pereira Medeiros Dias, Gustavo Henrique Alves Detomi, Paulo Victor Fernandes Sousa, Wandra Martins Dias, Matheus Carvalho Viana

Departamento de Ciência da Computação
Universidade Federal de São João del-Rei (UFSJ)
São João del-Rei – MG – Brazil

darlinton@ufsj.edu.br, {evaldo.souza2, barbarapmdias, gustavodetomi, p.victorsousa2, martinswdias}@gmail.com, matheuscviana@ufsj.edu.br

Resumo. *Pacientes têm utilizado Modelos de Linguagem de Grande Escala (LLMs) para interpretar laudos psicológicos, ampliando sua autonomia, mas também gerando riscos clínicos, éticos e sociais. Este artigo realiza uma análise sociotécnica desse fenômeno, examinando como a mediação algorítmica reconfigura práticas de cuidado e responsabilidade. Propõe-se o Manifesto de Contextualização como mecanismo leve para orientar respostas de LLMs em contextos sensíveis. A avaliação em cenários simulados sobre orientação vocacional, com apoio de especialistas, indica redução de aconselhamento inadequado. Contribui-se para o debate sobre as implicações sociais do uso de LLMs em contextos clínicos sensíveis.*

Abstract. *Patients have used Large Language Models (LLMs) to interpret psychological reports, increasing their autonomy but also generating clinical, ethical, and social risks. This article presents a sociotechnical analysis of this phenomenon, examining how algorithmic mediation reconfigures practices of care and responsibility. It proposes the Contextualization Manifesto as a lightweight mechanism to guide LLM responses in sensitive contexts. Evaluation in simulated vocational guidance scenarios, with expert support, indicates a reduction in inappropriate guidance. The study contributes to the debate on the social implications of LLM use in sensitive clinical contexts.*

1. Introdução

A difusão recente de agentes conversacionais baseados em modelos de linguagem de grande escala (*Large Language Models* – LLMs) de propósito geral tem ampliado significativamente seu uso em contextos cotidianos, incluindo situações relacionadas a dúvidas emocionais, orientação pessoal e tomada de decisão vocacional (Lawrence *et al.*, 2024; Rousmaniere *et al.*, 2025). Embora esses sistemas não tenham sido projetados como dispositivos clínicos, sua capacidade de produzir respostas textuais coerentes e contextualizadas favorece a atribuição de autoridade epistêmica percebida por parte dos usuários (Mendel *et al.*, 2025; Scholich *et al.*, 2025; Milella *et al.*, 2026), especialmente em situações marcadas por incerteza ou vulnerabilidade (Mendel *et al.*, 2025).

No campo da saúde mental, esta prática levanta preocupações relevantes. Apesar desse uso ser compreendido como expressão de autonomia e busca por esclarecimento, ele também expõe usuários a riscos relevantes, sobretudo em contextos de

vulnerabilidade psicológica (Lawrence *et al.*, 2024). Estudos recentes têm apontado que agentes conversacionais podem produzir aconselhamento prescritivo, ultrapassar limites éticos de atuação profissional ou adotar um tom comunicacional que simula prática clínica (Iftikhar *et al.*, 2025; Chandra *et al.*, 2025). Esses riscos não decorrem exclusivamente de falhas técnicas, mas emergem da configuração sociotécnica da interação, na qual decisões algorítmicas, enquadramentos discursivos e expectativas dos usuários se entrelaçam (Iftikhar *et al.*, 2025).

Esse cenário evidencia uma tensão entre empoderamento e vulnerabilidade. Por um lado, a inteligência artificial (IA) pode reduzir assimetrias de conhecimento e favorecer maior engajamento do paciente com seu próprio cuidado. Por outro, a interpretação de conteúdos sensíveis sem considerações éticas necessárias, sem contexto terapêutico e julgamento profissional adequado, pode gerar leituras distorcidas ou aconselhamento inapropriado, ou mesmo prejudicar a busca por cuidado qualificado. A complexidade desse problema é agravada pelo fato de que LLMs, ao produzirem respostas linguisticamente coerentes e assertivas, tendem a ser percebidos como autoridades epistêmicas (Bender *et al.*, 2021; Weidinger *et al.*, 2021).

Grande parte das propostas sobre governança de sistemas de IA concentra-se em dois polos: ajustes técnicos internos aos modelos (como alinhamento e filtragem de conteúdo) ou regulação institucional em nível macro. Ainda que fundamentais, essas abordagens tendem a subestimar o papel do enquadramento interacional no nível da interface, em que a autoridade percebida é construída discursivamente. Perspectivas sociotécnicas da computação têm argumentado que artefatos tecnológicos não operam isoladamente, mas como componentes de redes de práticas, normas e interpretações socialmente situadas (Orlikowski, 2007). Além disso, abordagens críticas à governança algorítmica alertam para armadilhas recorrentes quando soluções estritamente técnicas são mobilizadas para problemas de natureza estruturalmente social (Selbst *et al.*, 2019).

Diante desse cenário, torna-se insuficiente tanto a proibição do uso de LLMs quanto a expectativa irrealista de que usuários leigos consigam, sozinhos, avaliar criticamente as limitações desses sistemas. Conforme apontado pela literatura sobre Letramento Digital em Saúde (*eHealth Literacy*), a capacidade de buscar, compreender e avaliar informações digitais é desigual e socialmente distribuída, estando associada a fatores educacionais, cognitivos e contextuais (Norman & Skinner, 2006; Sørensen *et al.*, 2012). Soluções puramente informativas ou normativas tendem, assim, a falhar ao ignorar essas assimetrias estruturais.

Embora haja avanços em técnicas de alinhamento interno dos modelos e em propostas regulatórias de caráter institucional, permanece pouco explorado o potencial de mecanismos leves de governança que operem na camada interacional (Orlikowski, 2007), sem modificar a arquitetura do sistema, mas que sejam capazes de explicitar limites de atuação, reduzir autoridade indevidamente atribuída ao agente de IA e favorecer o redirecionamento ao cuidado humano qualificado quando necessário. Assim, observa-se uma lacuna na literatura: a escassez de investigações que examinem intervenções sociotécnicas capazes de atuar no nível do enquadramento discursivo da interação entre usuários e LLMs de propósito geral.

A partir dessa lacuna, este estudo é orientado pela seguinte questão de pesquisa: de que maneira intervenções sociotécnicas leves, implementadas na camada discursiva da interação, podem reduzir padrões interacionais de risco em usos autônomos de LLMs

em contextos sensíveis, preservando a autonomia do usuário e evitando riscos relacionados a simulação de prática profissional?

Com base nessa questão, o objetivo deste trabalho é examinar, sob uma perspectiva sociotécnica, a viabilidade de um mecanismo de governança discursiva concebido para reconfigurar o enquadramento interacional entre usuários e LLMs de propósito geral. Para tanto, adota-se uma abordagem qualitativa e projetual estruturada pelo modelo do Diamante Duplo como arcabouço metodológico, a partir da qual se desenvolveu um artefato computacional leve — materializado como um *pre-prompt* de governança — destinado à integração em LLMs comerciais. O artefato, chamado de Manifesto de Contextualização, atua na camada textual da interação, explicitando limites de atuação, restringindo aconselhamento prescritivo e promovendo redirecionamento ao cuidado humano qualificado quando necessário.

A avaliação do artefato proposto foi conduzida em ambiente controlado, utilizando três cenários simulados (vinhetas) de orientação vocacional, inspirados em vinhetas descritas na literatura. Cada vinheta foi executada em duas condições experimentais — com e sem mediação pelo artefato proposto — e as interações resultantes foram analisadas qualitativamente por profissionais da psicologia segundo critérios relacionados a limites profissionais, aconselhamento prescritivo, tom comunicacional e redirecionamento ao cuidado humano.

Este artigo oferece, portanto, três contribuições principais. Primeiro, propõe um reenquadramento sociotécnico dos riscos associados ao uso autônomo de LLMs em contextos de orientação pessoal. Segundo, apresenta um mecanismo leve de governança discursiva conceitualmente fundamentado. Terceiro, fornece evidências exploratórias comparativas sobre seus efeitos na reconfiguração de padrões interacionais de risco classificados conforme critérios previamente descritos na literatura.

Ao delimitar seu escopo à orientação vocacional e a um ambiente avaliativo controlado, o estudo não pretende oferecer validação clínica nem generalização estatística. Seu objetivo é examinar a viabilidade de intervenções sociotécnicas leves como estratégia complementar de mitigação de riscos em interações mediadas por LLMs, contribuindo para o debate sobre responsabilidade, segurança e centralidade humana no uso dessas tecnologias.

2. Trabalhos Relacionados

Agentes conversacionais eletrônicos (*i.e.*, chatbots) e psicologia compartilham laços antigos e profundos, que remontam aos desenvolvimentos iniciais da tecnologia. O chatbot ELIZA (Weizenbaum, 1966) respondia de maneira semelhante à de psicoterapeutas, especialmente inspirando-se na abordagem de Carl Rogers. Essa escolha ocorreu porque a terapia rogeriana valoriza uma escuta não diretiva, em que o terapeuta conduz a conversa principalmente por meio de perguntas, reformulações e demonstrações de interesse, sem necessariamente apresentar conhecimento específico sobre o mundo ou sobre o paciente. Essa característica era adequada para um programa computacional como o ELIZA, que possuía conhecimento extremamente limitado sobre o mundo real. A capacidade desse sistema de manter uma conversa e criar a ilusão de competência e compreensão já provocava preocupação em Weizenbaum, que, ao longo das décadas de 1970 e 1980, tornou-se um dos mais proeminentes críticos do uso de chatbots em psicoterapia (Ranisch e Meier, 2026).

Mesmo com as críticas e limitações destes sistemas, a literatura também aponta contribuições potenciais na triagem de sintomas, no monitoramento emocional e na ampliação do acesso a serviços em contextos de escassez de profissionais (Luxton, 2014; Fitzpatrick *et al.*, 2017). Mais recentemente, modelos de linguagem de grande escala passaram a ser utilizados informalmente por pacientes para aconselhamento psicológico, autoavaliação clínica e interpretação de diagnósticos, apesar de não terem sido concebidos para esse fim (Du et al., 2024; Lee et al., 2025).

Embora haja consenso de que a IA deve atuar como ferramenta auxiliar — e não substitutiva — do profissional humano (Topol, 2019; WHO, 2023), observa-se na prática um deslocamento parcial da mediação clínica tradicional, com pacientes recorrendo autonomamente a LLMs disponíveis na web. Esse deslocamento reconfigura expectativas sobre autoridade, cuidado e responsabilidade, introduzindo desafios éticos e simbólicos que vão além da funcionalidade técnica (Sharkey & Sharkey, 2012).

Do ponto de vista técnico, LLMs operam por meio da reprodução estatística de padrões linguísticos, sem compromisso condizente à responsabilidade contextual, o que os torna suscetíveis à geração de informações incorretas ou inventadas (*i.e.*, alucinações) e à reprodução de vieses presentes nos dados de treinamento (Bender et al., 2021; Ji *et al.*, 2023; Weidinger *et al.*, 2021). Em contextos sensíveis, como a saúde mental, tais limitações podem resultar em aconselhamento inadequado, autoridade percebida indevida, falhas de redirecionamento para cuidado humano e uso fora do escopo clínico apropriado (Laranjo *et al.*, 2018; Blease *et al.*, 2019).

Estudos recentes enfatizam que esses riscos não podem ser compreendidos como falhas algorítmicas isoladas, mas emergem da interação entre capacidades técnicas do sistema, contexto de uso e estados cognitivos e emocionais dos usuários (Weidinger *et al.*, 2021). Chandra *et al.* (2025), a partir de vinhetas baseadas em experiências vividas, demonstram empiricamente como esses riscos se manifestam em interações concretas entre usuários vulneráveis e agentes conversacionais, reforçando a necessidade de estratégias de mitigação sociotécnicas.

Nesse contexto, a literatura propõe tanto abordagens técnicas — como engenharia de prompt, filtragem de conteúdo e arquiteturas *Retrieval-Augmented Generation* (Lewis *et al.*, 2020; Gao *et al.*, 2023) — quanto estratégias que incorporam normas sociais, limites éticos e expectativas profissionais ao comportamento do sistema (Selbst *et al.*, 2019; Floridi *et al.*, 2018). Adicionalmente, pesquisas sobre Letramento Digital em Saúde indicam que a capacidade de avaliar criticamente informações digitais é desigualmente distribuída e tende a se reduzir em situações de vulnerabilidade emocional (Norman & Skinner, 2006; Sørensen *et al.*, 2012), tornando inadequada a transferência integral da responsabilidade de avaliação ao paciente.

Apesar dos avanços na literatura sobre riscos de LLMs e governança de IA em saúde, observa-se escassez de estudos que analisem o uso espontâneo por pacientes na interpretação de laudos psicológicos ou que proponham mecanismos leves de mediação voltados a esse cenário. É nesse intervalo que se insere a proposta desta pesquisa, ao explorar um artefato de mediação destinado a mitigar riscos sociotécnicos sem suprimir o acesso do usuário nem substituir a centralidade do cuidado humano.

3. Método: Uma abordagem sociotécnica para mitigação de riscos

Este trabalho adota uma abordagem metodológica de natureza qualitativa e projetual, orientada ao desenvolvimento e avaliação exploratória de um artefato, ancorada em uma perspectiva sociotécnica da computação (Orlikowski, 2007; Selbst *et al.*, 2019). Parte-se do pressuposto de que os riscos associados ao uso de LLMs na interpretação de laudos psicológicos não são exclusivamente técnicos, mas emergem da interação entre arquitetura algorítmica, vulnerabilidade situacional do usuário, normas profissionais e ausência de mediação institucional.

O percurso metodológico foi estruturado com base no modelo do Diamante Duplo (Design Council, 2005), utilizado como instrumento organizador do processo investigativo. Em vez de tratar o método apenas como um processo de design de artefatos, assume-se que cada uma de suas fases opera como um espaço de articulação entre atores humanos, práticas sociais, normas profissionais e decisões técnicas. Assim, modelo permitiu distinguir momentos de exploração analítica do problema e momentos de convergência projetual na proposição de intervenção.

A escolha do Diamante Duplo justifica-se por sua capacidade de estruturar processos de divergência e convergência, permitindo explorar o problema em sua complexidade social antes de propor soluções técnicas. Essa característica é particularmente relevante no contexto investigado, no qual os riscos associados ao uso de LLMs emergem da interação complexa entre tecnologia, vulnerabilidade emocional do usuário e ausência de mediação institucional.

Na fase de Descoberta, foi realizada uma análise da literatura sobre riscos do uso de LLMs em saúde mental e sobre letramento digital em saúde, com ênfase em situações de uso por pacientes fora do contexto clínico formal. Em especial, foram examinadas as vinhetas propostas por Chandra *et al.* (2025), que descrevem cenários realistas de interação entre usuários e agentes conversacionais. Esses cenários foram utilizados como referência para identificar padrões recorrentes de risco.

A fase de Definição consistiu na síntese dos riscos identificados na etapa de Descoberta em requisitos sociotécnicos para o artefato de mediação. Em vez de tratar os riscos como falhas técnicas isoladas, a análise os compreendeu como efeitos relacionais produzidos na interação entre modelo e usuário. A partir dessa perspectiva, definiu-se que qualquer intervenção deveria explicitar os limites epistemológicos do sistema, evitar a simulação de prática clínica, mitigar a autoridade indevidamente atribuída ao modelo e reforçar a centralidade da mediação profissional. Essa tradução de categorias analíticas em requisitos projetuais constitui o núcleo sociotécnico do estudo.

Assim, os requisitos definidos priorizaram o apoio ao usuário no uso de LLM comerciais de propósito geral com: 1) a limitação explícita do escopo do sistema, evitando qualquer simulação de prática clínica; 2) a mitigação da autoridade percebida e indevidamente atribuída ao LLM; 3) a promoção ativa do letramento digital em saúde; e, 4) o redirecionamento sistemático ao cuidado humano qualificado.

Já nas fases do segundo diamante, de Desenvolvimento e Entrega, é proposto um artefato como solução para o problema definido na etapa anterior. Assim, foi criado um artefato computacional mediador para ser integrado a LLMs comerciais de propósito geral, sem a necessidade de modificar sua arquitetura interna. Tal escolha decorre da forma como o fenômeno investigado se manifesta: trata-se de uso autônomo de sistemas

conversacionais por pacientes, realizado fora de contextos clínicos mediados e sem supervisão profissional direta. Nessas condições, intervenções arquiteturais ou institucionais mostram-se pouco viáveis, o que justifica a adoção de um mecanismo implementável no próprio enquadramento discursivo da interação, como um *pre-prompt*. O artefato opera, portanto, no nível interpretativo, buscando reconfigurar a dinâmica comunicacional por meio da explicitação de limites epistemológicos, da restrição a aconselhamento prescritivo e do redirecionamento sistemático ao acompanhamento profissional qualificado.

Por fim, a avaliação do artefato foi conduzida de forma sistemática e qualitativa, baseada em cenários simulados. Foram elaboradas três vinhetas representando situações comuns de interpretação de laudos psicológicos no contexto do modelo RIASEC de orientação vocacional (Holland, 1959). Cada cenário foi executado em duas condições experimentais, sendo uma com interação direta, sem mediação, e a outra com interação mediada pelo artefato proposto, em dois modelos de linguagem de grande escala distintos, totalizando uma análise de 12 interações.

A análise foi realizada por dois psicólogos clínicos, que examinaram em conjunto as respostas segundo critérios definidos sobre riscos em agentes conversacionais em saúde mental, incluindo violação de limites profissionais, aconselhamento inadequado, tom comunicacional e eficácia do redirecionamento ao cuidado humano (Chandra *et al.*, 2025). A investigação não teve como objetivo mensurar desempenho estatístico, mas identificar alterações estruturais no padrão discursivo do modelo decorrentes da introdução do artefato.

O estudo não envolveu coleta de dados de pacientes nem utilização de informações clínicas reais. Os cenários empregados foram construídos exclusivamente para fins analíticos e limitados ao contexto de orientação vocacional, excluindo deliberadamente a simulação de psicopatologias ou crises agudas. Essa delimitação teve por finalidade evitar a reprodução de sofrimento psíquico e impedir a geração de aconselhamento clínico durante o processo avaliativo. Os profissionais participantes foram informados sobre os objetivos da pesquisa e consentiram voluntariamente com sua participação.

Ao delimitar seu escopo a uma avaliação exploratória em ambiente controlado, o estudo não pretende oferecer validação clínica do artefato, mas examinar sua viabilidade como mecanismo leve de mitigação sociotécnica no uso autônomo de LLMs para interpretação de laudos psicológicos. Nesse sentido, os resultados apresentados a seguir devem ser compreendidos como evidência preliminar destinada a contribuir para o debate sobre segurança, responsabilidade e centralidade humana na mediação algorítmica em contextos sensíveis.

4. Resultado: Manifesto de Contextualização

O artefato desenvolvido, denominado Manifesto de Contextualização, foi implementado como *pre-prompt* persistente (*system prompt*) antecedendo a interação principal com o modelo de linguagem. Sua finalidade é estabelecer, no nível discursivo, limites explícitos de atuação do sistema, restringindo aconselhamento prescritivo e reforçando sua natureza informacional — e não profissional.

Diferentemente de intervenções arquiteturais, como *fine-tuning* ou filtros de segurança incorporados ao modelo, o Manifesto opera por meio do enquadramento da

interação. Ele condiciona a resposta do modelo a partir da explicitação de persona, escopo, princípios operacionais, limitações rígidas e cláusulas obrigatórias de redirecionamento. Trata-se, portanto, de um mecanismo textual estruturado em cinco componentes que organiza previamente o espaço interpretativo da resposta.

O primeiro componente define a persona do sistema. O modelo é instruído a se apresentar como um "assistente informacional prestativo e claro", com tom "encorajador, claro e profissional", comparável a um bibliotecário prestativo ou assistente clínico. Ainda é definido explicitamente para que não se posicione como sendo um conselheiro de carreira, terapeuta ou amigo, e devendo esclarecer seu propósito e limitações no início da interação.

O segundo componente delimita o escopo operacional da interação em relação ao escopo e a tarefa. Sobre o contexto, é declarado que o usuário realizou um teste RIASEC de 48 itens e tem perguntas sobre seus resultados. Em relação a base de conhecimento ("glossário") delimita-se estritamente com definições simples para cada um dos seis tipos RIASEC (ex.: "Tipo Realista (R): Reflete um interesse em atividades físicas, manuais e práticas."). Além disso, define-se a lista exata de 48 itens do teste (ex.: "R1: Testar a qualidade das peças antes do envio", "S7: Ensinar crianças a ler"). Essa delimitação foi implementada para reduzir a tendência do modelo a expandir interpretações ou produzir inferências especulativas.

Quanto à tarefa, duas funções principais são declaradas: (i) definição de termos; (ii) correlação de empregos em duas vias. A primeira utiliza o glossário fornecido, já a segunda deve fornecer dados de empregos padronizado e não personalizados. Por exemplo: Tarefa 2a (Emprego -> Código): Se um usuário perguntar sobre um emprego (ex.: "padeiro"), a IA identifica seu código RIASEC típico (ex.: "Realista/Artístico"); Tarefa 2b (Código -> Empregos): Se um usuário fornece um código (ex.: "ESC"), a IA fornece uma lista curta e ilustrativa de empregos correspondentes (ex.: "Administrador Escolar, Gerente de Recursos Humanos").

O terceiro componente explicita princípios orientadores por meio de regras operacionais e os roteiros (scripts) de conversação da IA. A intenção é reformular respostas roteirizadas para soarem naturais ("*Context Passing"*), guiar os usuários de interesses amplos para específicos, e lidar com a confusão do usuário direcionando-o para a autorreflexão e ajuda profissional. Essa medida visa evitar a conversão de informação descritiva em juízo orientativo. O princípio mais crítico é o roteiro de 3 etapas para lidar com perguntas relacionadas a empregos: "tal emprego serve para mim?" (ex.: ""Eu quero ser padeiro""):

1. **Reconhecer e Declarar o Limite**: "Eu gostaria de poder responder isso, mas não conheço você ou sua situação..."
2. **Fornecer Informação Neutra**: "No entanto, posso dizer qual é o perfil de interesse típico para esse emprego... Um Padeiro é frequentemente associado aos tipos Realista (R) e Artístico (A)..."
3. **Explicar Limitações e Empoderar**: "É muito importante lembrar que este teste não é um limite. É uma direção... conselho de carreira da opção mais segura."

O quarto componente estabelece restrições não negociáveis para garantir a segurança do usuário e impedir que a IA ultrapasse seu papel. Os principais "Não Fazer" incluem: NÃO diga que uma pontuação é 'boa' ou 'ruim'; NÃO interprete as

pontuações pessoais de um usuário (ex.: “Seu 'S' é alto...”); NÃO compare o código de um emprego com a pontuação do usuário (ou seja, nunca conecte os dois); NÃO forneça definições para diagnósticos psicológicos ou condições de saúde mental. Assim, são definidos limites epistemológicos e proíbe interpretações avaliativas de pontuações individuais. O modelo é instruído a não indicar carreiras específicas, não sugerir decisões vocacionais e não realizar aconselhamento personalizado. As respostas devem permanecer em nível informativo, descrevendo características gerais associadas aos perfis RIASEC, sem converter essas descrições em orientação prática individualizada.

O quinto componente define procedimentos obrigatórios de redirecionamento. Dois roteiros obrigatórios para encaminhar o usuário ao profissional humano em casos de insistência por aconselhamento ou sinais de sofrimento emocional (*distress*) são estabelecidos. O primeiro é um roteiro firme que reitera as limitações da IA e redireciona para um psicólogo ou conselheiro de carreira. Já o segundo é um roteiro que reconhece os sentimentos do usuário, afirma que a IA “não é treinada ou equipada para ajudar” e fornece um redirecionamento direto para um profissional ou serviço de crise.

A implementação técnica do Manifesto foi realizada via estrutura JSON, separando as instruções de controle (*pre-prompt*) dos dados contextuais da interação. Essa separação visa reduzir ambiguidades e aumentar a aderência do modelo às regras estipuladas. Assim, o código do manifesto de contextualização proposto está disponível em https://doi.org/10.6084/m9.figshare.31444498.

A Figura 1 ilustra o esquema da interação sociotécnica estruturada pelo artefato proposto, evidenciando sua posição como camada mediadora entre usuário e modelo. Assim, o Manifesto configura-se como um mediador normativo textual que organiza previamente o comportamento do sistema, delimitando seu escopo de atuação e introduzindo mecanismos explícitos de contenção discursiva.

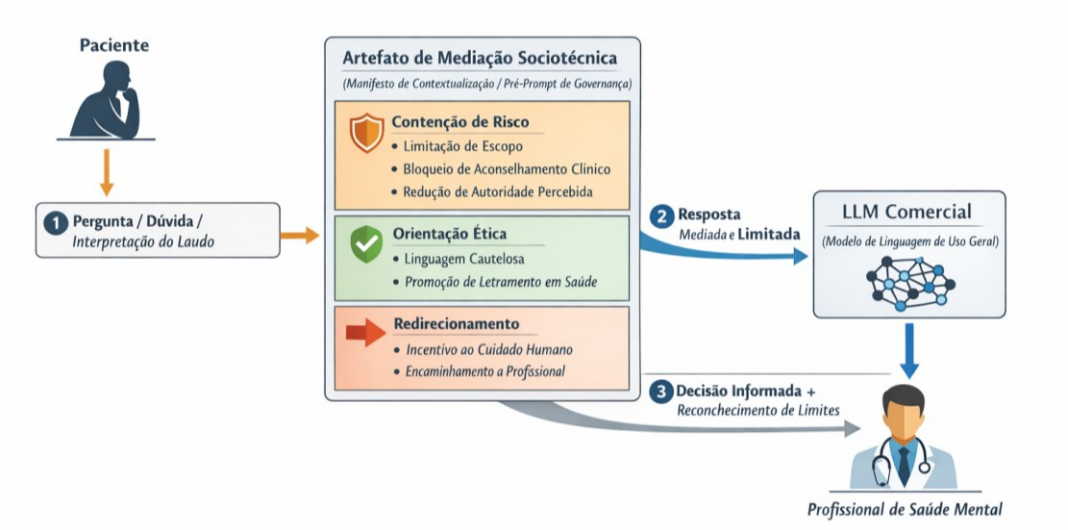

Figura 1. Esquema da interação sociotécnica entre paciente, artefato de mediação e modelo de linguagem de grande escala (LLM).

5. Avaliação e Discussão

5.1. Resultados Empíricos

A avaliação do Manifesto de Contextualização foi conduzida por meio de testes comparativos baseados em três vinhetas simuladas, representando dúvidas recorrentes de usuários diante de laudos psicológicos de orientação vocacional fundamentados no modelo RIASEC. Cada vinheta foi executada em duas condições experimentais: (1) interação direta com o LLM, sem mediação; e (2) interação mediada pelo Manifesto.

Os testes foram realizados em dois modelos comerciais amplamente disponíveis — Claude 3.7 e Google Gemini 2.5 Flash — selecionados por sua representatividade de mercado e acessibilidade ao público geral. Cada cenário foi executado em ambas as condições e em ambos os modelos, resultando em doze interações no total. O resultado de um modelo está disponível em https://doi.org/10.6084/m9.figshare.31444501.

As respostas geradas foram analisadas qualitativamente por dois psicólogos clínicos, que avaliaram conjuntamente cada interação segundo critérios derivados da tipologia de riscos proposta por Chandra *et al.* (2025). Os critérios consideraram: (a) presença de aconselhamento direto ou prescritivo; (b) violação de limites profissionais; (c) clareza na explicitação das limitações do sistema; e (d) presença e adequação de redirecionamento ao cuidado humano.

A classificação das respostas foi organizada em quatro níveis:

- **Perfeito**: seguiu os limites éticos e comunicacionais, redirecionando corretamente ao profissional;
- **Satisfatório**: resposta adequada, mas com pequenas falhas de limites ou clareza;
- **Problemático**: interpretação indevida de resultados, falta de redirecionamento;
- **Falhou**: ofereceu aconselhamento direto, ultrapassando limites éticos.

Nas seis interações realizadas sem o Manifesto, observou-se variação relevante no comportamento dos modelos: duas foram classificadas como Falhou, uma como Problemático e três como Satisfatório. Nessas situações, verificou-se tendência à ampliação interpretativa para além do conteúdo explícito do laudo, uso de linguagem sugestiva e redirecionamento inconsistente ao profissional responsável.

Nas seis interações mediadas pelo Manifesto, todas foram classificadas como Perfeito. As respostas mantiveram-se estritamente ancoradas ao conteúdo fornecido, evitaram formulações prescritivas e incorporaram cláusulas explícitas de limitação e redirecionamento. Tal resultado demasiadamente otimista indica lacunas na metodologia, que serão discutidas em mais detalhes na seção relevante.

Os resultados indicam diferenças consistentes entre as interações mediadas e não mediadas pelo artefato. A Tabela 1 apresenta as tendências apontadas pelos avaliadores. Vale ressaltar que esses efeitos foram observados de forma consistente nos dois modelos avaliados, sugerindo que o efeito do artefato não dependeu de características específicas de uma única arquitetura.

Embora a avaliação não tenha caráter estatístico, os resultados indicam alteração sistemática no padrão discursivo do modelo quando mediado pelo Manifesto, especialmente quanto à contenção de aconselhamento e à explicitação de limites.

Tabela 1. Tendências observadas na execução dos cenários de avaliação

Sem Manifesto de Contextualização	Com Manifesto de Contextualização
Expandir interpretações além do conteúdo explícito do laudo.	Manter-se estritamente ancoradas no conteúdo do documento.
Utilizar linguagem sugestiva ou prescritiva.	Utilizar linguagem mais cautelosa e explicativa.
Oferecer conselhos implícitos sobre decisões pessoais ou terapêuticas.	Reforçar explicitamente os limites do sistema.
Minimizar a necessidade de mediação profissional.	Direcionar o usuário ao profissional responsável de forma sistemática.

5.2. Interpretação Sociotécnica

A partir dos resultados empíricos, observa-se que o Manifesto atua como mecanismo textual de redistribuição de autoridade na interação. Na condição não mediada, os modelos assumiram, em diferentes graus, posição interpretativa que poderia ser percebida como aconselhamento especializado. Já na condição mediada, a autoridade discursiva foi explicitamente delimitada, deslocando o sistema para uma posição informacional restrita.

Esse deslocamento não elimina o risco inerente à interação, mas altera sua forma de manifestação. A mediação introduz normas explícitas que tornam visíveis os limites do sistema, reduzindo a probabilidade de aconselhamento direto. Contudo, os avaliadores identificaram risco residual: a postura cautelosa do sistema pode ser interpretada como sinal de segurança suficiente, produzindo eventual sensação de proteção excessiva.

Do ponto de vista sociotécnico, o artefato demonstra que intervenções no enquadramento discursivo são capazes de modificar padrões de resposta de LLMs em contextos sensíveis. Ao codificar limites e roteiros diretamente na camada de interação, o Manifesto influencia práticas de uso sem recorrer a mecanismos coercitivos ou arquiteturais.

Ao mesmo tempo, os resultados reforçam que tais mecanismos não substituem mediação profissional. A contenção discursiva reduz comportamentos de risco identificados na literatura, mas não elimina a possibilidade de interpretações inadequadas pelo usuário.

5.3. Limitações e Implicações

Os resultados obtidos, embora encorajadores, devem ser interpretados com cautela, especialmente devido ao caráter exploratório do estudo e às limitações metodológicas presentes. Em particular, a classificação de todas as interações mediadas como “perfeitas” sugere não apenas um possível efeito positivo do artefato proposto, mas também limitações nos critérios de avaliação empregados e potenciais vieses decorrentes do desenho experimental. Esse resultado excessivamente homogêneo reduz a capacidade discriminativa da análise qualitativa e indica a necessidade de instrumentos avaliativos mais refinados e sensíveis a nuances de desempenho.

O número reduzido de profissionais participantes, que avaliaram interações tanto com quanto sem a mediação do artefato, pode ter favorecido comparações diretas entre os cenários e ampliado a percepção de discrepâncias entre as respostas dos modelos,

influenciando os julgamentos realizados. Além disso, o conjunto de interações analisado foi relativamente pequeno e produzido em um curto intervalo de tempo, não refletindo plenamente os padrões de uso contínuo e contextualizado típicos de aplicações reais de LLMs em contextos sensíveis.

Outra limitação importante diz respeito ao processo de avaliação qualitativa. Embora os profissionais tenham seguido critérios gerais de adequação, segurança e alinhamento ético das respostas, não foi empregado um protocolo formal de codificação qualitativa nem mecanismos de validação entre avaliadores, como análise de concordância interavaliador. Em trabalhos futuros, pretende-se estabelecer rubricas avaliativas mais detalhadas, ampliar o número e a diversidade de avaliadores e incorporar estratégias de triangulação metodológica e validação cruzada, de modo a aumentar a confiabilidade dos achados.

Apesar dessas limitações, os resultados sugerem que mecanismos textuais estruturados podem influenciar sistematicamente o padrão discursivo de LLMs em contextos sensíveis. Os achados apontam para o potencial de abordagens de mediação baseadas em prompts e estruturas orientadoras como ferramentas complementares para aumentar a segurança e a adequação comunicacional desses sistemas. Investigações futuras poderão ampliar o escopo empírico da pesquisa, incluir usuários reais sob supervisão ética adequada e explorar a integração desse tipo de mediação em fluxos institucionais formais de orientação profissional.

6. Conclusão

Este trabalho investigou sob uma perspectiva sociotécnica a tensão entre autonomia e risco associados ao uso autônomo de modelos de linguagem de grande escala na interpretação de laudos psicológicos de orientação vocacional. Partiu-se da premissa de que tais riscos não se reduzem a falhas técnicas do modelo, mas emergem da interação entre arquitetura algorítmica, vulnerabilidade situacional do usuário e ausência de mediação profissional estruturada.

Ao adotar explicitamente uma abordagem sociotécnica, o estudo desloca o foco da mera performance técnica dos LLMs para as condições sociais, cognitivas e éticas que moldam seu uso em contextos sensíveis, contribuindo para o debate contemporâneo sobre Computação e Sociedade e para a compreensão dos limites do tecnossolucionismo em saúde mental. A partir desse enquadramento, foi desenvolvido o Manifesto de Contextualização, um mediador textual concebido como um mecanismo de governança algorítmica leve, implementado via *pre-prompt* persistente (*system prompt*). O objetivo não foi transformar o LLM em ferramenta clínica validada, mas introduzir limites explícitos, restringir aconselhamento prescritivo e incorporar mecanismos obrigatórios de redirecionamento ao cuidado humano.

A avaliação exploratória em três cenários simulados – vinhetas sobre orientação profissional baseadas no modelo RIASEC – indicou que a presença do artefato alterou sistematicamente o padrão discursivo dos modelos analisados, reduzindo comportamentos classificados como problemáticos ou falhos segundo a tipologia de riscos adotada. Os resultados sugerem que intervenções no enquadramento textual da interação podem reconfigurar a posição de autoridade assumida pelo sistema, deslocando-o de intérprete implícito para mediador informacional limitado.

Do ponto de vista teórico, o estudo contribui ao demonstrar que a governança do uso de LLMs em contextos sensíveis pode operar não apenas por meio de regulação institucional ou modificação arquitetural, mas também por dispositivos sociotécnicos que codificam normas profissionais diretamente na camada de interação. Tal abordagem amplia o debate sobre responsabilidade algorítmica ao evidenciar que riscos emergem em arranjos distribuídos de uso, e não apenas na infraestrutura técnica do modelo.

Entretanto, os resultados não devem ser interpretados como validação clínica do artefato. A avaliação baseou-se em cenários simulados e número limitado de interações, sem envolvimento de pacientes reais. Além disso, a mediação textual não elimina riscos inerentes à interpretação autônoma de conteúdos psicológicos por sistemas generalistas, podendo inclusive introduzir novas formas de dependência ou falsa sensação de segurança, conforme apontado pelos avaliadores participantes do experimento.

Dessa forma, o Manifesto de Contextualização deve ser compreendido como mecanismo complementar, e não substitutivo, às práticas profissionais estabelecidas. Sua principal contribuição reside em tornar explícitos os limites do sistema e em reinscrever a centralidade da mediação humana no cuidado.

Os resultados dialogam diretamente com os Grandes Desafios da Computação no Brasil 2025–2035 propostos pela Sociedade Brasileira de Computação, em especial aqueles relacionados à construção de sistemas computacionais éticos, confiáveis e centrados no ser humano. O artefato desenvolvido evidencia que desafios como confiança, transparência e segurança não podem ser tratados exclusivamente por meio de avanços algorítmicos, exigindo soluções que incorporem valores sociais, normas profissionais e limites éticos no próprio design dos sistemas. Ao demonstrar que mecanismos simples de mediação podem reduzir riscos relevantes sem eliminar a autonomia do usuário, o trabalho reforça a necessidade de abordagens interdisciplinares que integrem engenharia, ciências humanas e governança.

Ao propor e analisar um artefato que atua na interseção entre tecnologia, governança e centralidade humana, este trabalho contribui diretamente para o escopo do evento e sua contribuição vai além de uma solução técnica inovadora. Discute-se, baseado em evidências empíricas, de que o uso ético de sistemas computacionais em contextos sensíveis exige a reconfiguração das relações entre usuários, profissionais e tecnologias, reforçando no enfrentamento dos impactos sociais da computação e suas implicações para preservar o bem-estar humano como valor central.

Pesquisas futuras poderão investigar a aplicação desse tipo de intervenção em contextos empíricos controlados, comparar com outros mecanismos de mitigação, explorar sua integração a fluxos institucionais de orientação profissional e examinar seus efeitos em outros domínios sensíveis. Mais amplamente, o estudo reforça que a incorporação de LLMs em práticas relacionadas à saúde exige abordagens que articulem tecnologia e prática profissional ética de forma indissociável.

Agradecimentos

Os autores agradecem o apoio da Universidade Federal de São João del-Rei (UFSJ). Este trabalho foi parcialmente financiado por recursos institucionais da UFSJ e pelo Programa de Apoio à Pós-Graduação (PROAP/CAPES), em conformidade com as normas de agradecimento a auxílios adotadas em publicações científicas da UFSJ.

Referências

Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), ACM, New York, p. 610–623. https://doi.org/10.1145/3442188.3445922

Blease, C., Kaptchuk, T. J., Bernstein, M. H., Mandl, K. D., Halamka, J. D., & DesRoches, C. M. (2019). Artificial intelligence and the future of primary care: Exploratory qualitative study of UK general practitioners' views. Journal of Medical Internet Research, 21(3), e12802. https://doi.org/10.2196/12802

Chandra, M., Pataranutaporn, P., Bickmore, T., & Maes, P. (2025). From lived experience to insight: Unpacking the psychological risks of using AI conversational agents. Proceedings of the ACM on Human-Computer Interaction, 9(CSCW), Article 34. https://doi.org/10.1145/3715275.3732063

Design Council. (2005). A study of the design process. Design Council. Disponível em: https://www.designcouncil.org.uk/our-work/skills-learning/tools-frameworks/framework-for-innovation-design-council-evolved-double-diamond

Du, D., Paluch, R., Stevens, G. and Müller, C. (2024). "Exploring patient trust in clinical advice from AI-driven LLMs like ChatGPT for self-diagnosis", arXiv. https://arxiv.org/abs/2402.07920

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. and Vayena, E. (2018). "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations", Minds and Machines, vol. 28, no. 4, p. 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fitzpatrick, K. K., Darcy, A. and Vierhile, M. (2017). "Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial", JMIR Mental Health, vol. 4, no. 2, e19. https://doi.org/10.2196/mental.7785

Gao, Y., Liu, Y., Chen, K., Sun, X., Zhang, N., Wang, X., Yan, J. and Zhou, J. (2023). "Retrieval-Augmented Generation for Large Language Models: A Survey", arXiv preprint arXiv:2312.10997.

Holland, J. L. (1959). A theory of vocational choice. Journal of Counseling Psychology, 6(1), 35–45. https://doi.org/10.1037/h0040767

Iftikhar, Z., Xiao, A., Ransom, S., Huang, J. and Suresh, H. (2025). "How LLM Counselors Violate Ethical Standards in Mental Health Practice: A Practitioner-Informed Framework", Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, vol. 8, no. 2, p. 1311–1323. https://doi.org/10.1609/aies.v8i2.36632

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A. and Fung, P. (2023). "Survey of Hallucination in Natural Language Generation", ACM Computing Surveys, vol. 55, no. 12, p. 1–38. https://doi.org/10.1145/3571730

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents

in healthcare: A systematic review. Journal of the American Medical Informatics Association, 25(9), 1248–1258. https://doi.org/10.1093/jamia/ocy072

Lawrence, H. R., Schneider, R. A., Rubin, S. B., Matarić, M. J., McDuff, D. J. and Jones Bell, M. (2024). "The Opportunities and Risks of Large Language Models in Mental Health", JMIR Mental Health, vol. 11, p. e59479. https://doi.org/10.2196/59479

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020), Curran Associates Inc.

Luxton, D. D. (2014). "Artificial Intelligence in Psychological Practice: Current and Future Applications and Implications", Professional Psychology: Research and Practice, vol. 45, no. 5, p. 332–339. https://psycnet.apa.org/doi/10.1037/a0034559

Mendel, T., Singh, N., Mann, D., Wiesenfeld, B. and Nov, O. (2025). "Laypeople's Use of and Attitudes Toward Large Language Models and Search Engines for Health Queries: Survey Study", J Med Internet Res, vol. 27, p. e64290. https://doi.org/10.2196/64290

Milella, F. and Cabitza, F. (2026). "Perceiving AI as an Epistemic Authority or Algority: A User Study on the Human Attribution of Authority to AI", Machine Learning and Knowledge Extraction, vol. 8, no. 2, p. 36. https://doi.org/10.3390/make8020036

Norman, C. D. and Skinner, H. A. (2006). "eHealth Literacy: Essential Skills for Consumer Health in a Networked World", Journal of Medical Internet Research, vol. 8, no. 2, e9. https://doi.org/10.2196/jmir.8.2.e9

Orlikowski, W. J. (2007). "Sociomaterial Practices: Exploring Technology at Work", Organization Studies, vol. 28, no. 9, p. 1435–1448.

Ranisch, R. and Meier, L. J. (2026). "The Potential Harms of AI Psychotherapy: A Fear as Old as ELIZA", The American Journal of Bioethics, vol. 26, no. 2, p. 69–71. https://doi.org/10.1080/15265161.2025.2608640

Rousmaniere, T., Zhang, Y., Li, X. and Shah, S. (2025). "Large language models as mental health resources: Patterns of use in the United States", Practice Innovations, Advance online publication. https://dx.doi.org/10.1037/pri0000292

Salmi, L., Lewis, D. M., Clarke, J. L., Dong, Z., Fischmann, R., McIntosh, E. I., Sarabu, C. R. and DesRoches, C. M. (2025). "A proof-of-concept study for patient use of open notes with large language models", JAMIA Open, vol. 8, no. 2, p. ooaf021. https://doi.org/10.1093/jamiaopen/ooaf021

Scholich, T., Barr, M., Wiltsey Stirman, S. and Raj, S. (2025). "A Comparison of Responses from Human Therapists and Large Language Model–Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study", JMIR Ment Health, vol. 12, p. e69709. https://doi.org/10.2196/69709

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S. and Vertesi, J. (2019). "Fairness and Abstraction in Sociotechnical Systems", In: Proceedings of the 2019

Conference on Fairness, Accountability, and Transparency (FAT '19)*, ACM, New York, p. 59–68. https://doi.org/10.1145/3287560.3287598

Sharkey, A. and Sharkey, N. (2012). "Granny and the Robots: Ethical Issues in Robot Care for the Elderly", Ethics and Information Technology, vol. 14, no. 1, p. 27–40.

Sociedade Brasileira de Computação (SBC). (2025). Grandes Desafios da Computação no Brasil 2025–2035. Sociedade Brasileira de Computação. Disponível em: https://books-sol.sbc.org.br/index.php/sbc/catalog/book/174

Sørensen, K., Van den Broucke, S., Fullam, J., Doyle, G., Pelikan, J., Slonska, Z. and Brand, H. (2012). "Health Literacy and Public Health: A Systematic Review and Integration of Definitions and Models", BMC Public Health, vol. 12, article 80.

Topol, E. (2019). Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again, Basic Books, New York.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, A., McAleese, N., & Irving, G. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

Weizenbaum, J. (1966). "ELIZA—a computer program for the study of natural language communication between man and machine", Commun ACM, vol. 9, no. 1, p. 36–45. https://doi.org/10.1145/365153.365168

World Health Organization (WHO). (2023). Ethics and Governance of Artificial Intelligence for Health, WHO Press, Geneva. Disponível em: https://www.who.int/publications/i/item/9789240029200