Mediação de Privacidade para Interações com Modelos de Linguagem de Grande Porte

Felipe Diego Lobato da Silva[1],**Thiago Adriano Coleti**[1,2]

[1]Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP)
São Paulo – SP – Brasil

[2]Centro de Ciências Tecnológicas da Universidade Estadual do Norte do Paraná (UENP)
Bandeirantes, PR – Brasil

felipediegolobatodasilva@usp.br, thiago.coleti@uenp.edu.br

Abstract. *This paper presents a privacy mediation tool for interactions with large language models (LLM), implemented as a browser extension. The tool identifies personal and sensitive data in the text entered by the user before it is sent to the model, using pattern-based rules and Named Entity Recognition (NER) techniques. When sensitive information is detected, just-in-time visual alerts are displayed, allowing the user to decide whether to proceed, edit, or cancel the message submission.*

Resumo. *Este trabalho apresenta uma ferramenta de mediação de privacidade para interações com modelos de linguagem de grande porte (LLM), implementada como uma extensão de navegador. A ferramenta identifica dados pessoais e sensíveis no texto inserido pelo usuário antes do envio ao modelo, utilizando regras baseadas em padrões e técnicas de reconhecimento de entidades nomeadas (NER). Quando informações sensíveis são detectadas, alertas visuais just-in-time são exibidos, permitindo que o usuário decida se deseja prosseguir, editar ou cancelar o envio da mensagem.*

1. Introdução

Aplicações conversacionais são ferramentas que simulam a interação humana por meio de linguagem natural em diversas plataformas (Lima et al., 2025). Essas ferramentas utilizam Modelos de Linguagem de Grande Porte (*Large Language Models* – LLM), e tornaram-se comuns para apoiar tarefas como atendimento, pesquisas, cuidados entre outras (Naveed et al., 2025). As LLM ampliam as possibilidades de uso das aplicações, mas podem incentivar os usuários a compartilhar dados pessoais e sensíveis, em razão da utilidade percebida e da natureza humanizada da interação (Li et al., 2024).

Interfaces conversacionais baseadas em LLMs favorecem interações contínuas e linguisticamente naturais, aproximando-se de padrões típicos da comunicação humana. Em sistemas dessa natureza, características antropomórficas presentes na interface e no comportamento conversacional podem influenciar a percepção e o comportamento dos usuários durante a interação, incluindo aspectos relacionados ao self-disclosure e ao compartilhamento de informações pessoais (Seeger et al., 2021).

O compartilhamento pode ocorrer de forma não intencional, devido à falta de conhecimento do usuário, ou pelo emprego de padrões enganosos (Baroni and Pereira,

2024). Apesar dos avanços recentes, a proteção de dados em aplicações conversacionais baseadas em LLM ainda enfrenta limitações associadas à complexidade das Políticas de Privacidade e Segurança (PPS) e à ausência de mecanismos de apoio ao usuário durante as interações (Chen et al., 2025; Freiberger et al., 2025a). Esses fatores podem dificultar a compreensão dos riscos envolvidos, ampliar riscos à privacidade e à proteção de dados e expor indevidamente o usuário (Leschanowsky et al., 2024).

Diante desse problema, este trabalho propõe uma aplicação de mediação de privacidade para interações com aplicações conversacionais baseadas em LLM, voltada à identificação automática de dados pessoais potencialmente sensíveis, inseridos pelo usuário em interfaces conversacionais. Para isso, são empregadas técnicas de processamento de linguagem natural, com destaque para o Reconhecimento de Entidades Nomeadas (*Named Entity Recognition* – NER), técnica voltada à identificação e classificação automática de entidades relevantes em textos, como nomes de pessoas, organizações e localizações, favorecendo a extração estruturada de informações textuais (Keraghel and Nadif, 2025).

A ferramenta proposta busca apoiar o usuário durante a interação, auxiliando a tomada de decisão sobre o compartilhamento de dados pessoais por meio da identificação automática de informações potencialmente sensíveis e da sinalização de riscos de privacidade. Neste trabalho, o termo mediação de privacidade designa uma camada de apoio contextual à decisão do usuário, sem envolver bloqueio automático, gerenciamento formal de consentimento ou aplicação centralizada de políticas de privacidade.

2. Fundamentação teórica e Trabalhos correlatos

A proteção de dados em aplicações conversacionais baseadas em LLM não se restringe à existência de bases legais para o tratamento das informações, mas envolve também mecanismos que favoreçam transparência e controle informacional ao usuário (European Data Protection Supervisor, 2023). Em aplicações dessa natureza, a troca de dados ocorre de forma textual, dinâmica e, por vezes, pouco refletida, o que amplia a relevância de abordagens capazes de tornar mais visíveis os riscos associados ao compartilhamento de informações pessoais (Zhang et al., 2024).

Entre as abordagens técnicas que podem contribuir para esse apoio, destaca-se o Reconhecimento de Entidades Nomeadas (NER, do inglês, *Named Entity Recognition*), técnica voltada à identificação e classificação automática de entidades em textos, como nomes de pessoas, organizações e localizações (Yadav and Bethard, 2018). Em interações textuais com agentes conversacionais, essa técnica pode ser estendida à detecção de informações pessoais em diálogos, favorecendo o reconhecimento de elementos que demandam tratamento diferenciado quando associados a conteúdos sensíveis (Mina et al., 2024).

Na literatura recente é possível identificar trabalhos que investigam diferentes estratégias para apoiar a privacidade em aplicações conversacionais baseadas em LLM, incluindo mecanismos de apoio contextual ao usuário, minimização de dados e interpretação de políticas de privacidade.

Chen et al. (2025) apresentaram o CLEAR, uma extensão de navegador voltada à análise contextual de políticas de privacidade em aplicações baseadas em LLM. A fer-

ramenta auxilia usuários a compreender como suas informações podem ser tratadas, exibindo trechos relevantes das políticas e riscos potenciais no momento da interação. O estudo, baseado em workshops de co-design com 16 participantes, indicou aumento da conscientização sobre riscos de privacidade.

Zhou et al. (2025) desenvolveram o Rescriber, uma extensão de navegador voltada à minimização de dados em interações com chatbots baseados em LLM. A solução utiliza modelos de linguagem para detectar, destacar e sanitizar informações pessoais antes do envio da mensagem, permitindo ao usuário substituir ou abstrair trechos sensíveis. Avaliada com 12 usuários do ChatGPT, a ferramenta apresentou resultados positivos na redução do compartilhamento desnecessário de dados.

Freiberger et al. (2025b) propuseram o PRISMe, uma ferramenta interativa baseada em LLM para avaliação de políticas de privacidade. A proposta combina painel visual, explicações em linguagem natural e interface conversacional para auxiliar usuários na compreensão de documentos extensos e tecnicamente complexos. O estudo com 22 participantes indicou maior entendimento sobre o tratamento de dados pessoais e riscos de privacidade.

Ischen et al. (2020) investigaram preocupações de privacidade em interações com chatbots, analisando como características da interface e do comportamento conversacional podem influenciar a percepção dos usuários. Embora o estudo não trate de aplicações baseadas em LLM, os autores mostraram que elementos antropomórficos e contextos de interação mais humanizados podem aumentar a confiança e, consequentemente, estimular maior compartilhamento de informações pessoais. Os resultados evidenciam a importância de mecanismos que apoiem o usuário durante a interação com agentes conversacionais.

Os trabalhos apresentados evidenciam o avanço de soluções voltadas à proteção da privacidade em interações com modelos de linguagem de grande porte, seja por meio da análise de políticas de privacidade, da minimização de dados ou da oferta de explicações ao usuário durante a interação. A solução proposta neste artigo contribui ao introduzir um mecanismo de mediação de privacidade integrado diretamente ao fluxo de uso de interfaces conversacionais, atuando antes do envio da mensagem ao modelo, sem o uso de LLM em seu *background*.

3. Proposta de ferramenta de mediação

Esta seção apresenta informações sobre a ferramenta proposta, que tem por objetivo identificar possíveis compartilhamentos de dados pessoais em *chatbots* e orientar quanto aos riscos e ações de forma simples e compreensível.

3.1. Arquitetura da ferramenta

A arquitetura da ferramenta é composta por três elementos principais: (i) o módulo de verificação de segurança; (ii) o componente mediador; e (iii) o modelo de linguagem. A interação do usuário ocorre na interface conversacional web, utilizada como ponto de entrada para inserção de texto, mas que não integra diretamente a arquitetura interna da aplicação.

Quando o usuário aciona o envio da mensagem, a extensão intercepta o evento associado ao campo de entrada antes que o conteúdo seja encaminhado ao chatbot. O texto

é extraído do DOM da página e encaminhado ao módulo de verificação de segurança, no qual são aplicadas regras regex para padrões estruturados, como CPF, e-mail, números telefônicos e cartões bancários, e técnicas de NER para informações contextuais, como nomes próprios, organizações, localidades e datas. Os resultados são submetidos a validações contextuais e classificados heuristicamente conforme o potencial de sensibilidade e risco das informações detectadas.

Os resultados dessa análise são encaminhados ao componente mediador, responsável por apresentar alertas *just-in-time* ao usuário, contendo informações sobre o tipo de dado detectado, o nível de risco associado e as ações disponíveis.

Nessa organização, o mediador estabelece uma camada intermediária entre a interface conversacional e o modelo de linguagem, oferecendo suporte à tomada de decisão antes do envio da mensagem. Na Figura 1 é apresentada uma visão geral dos componentes da arquitetura proposta e de suas relações.

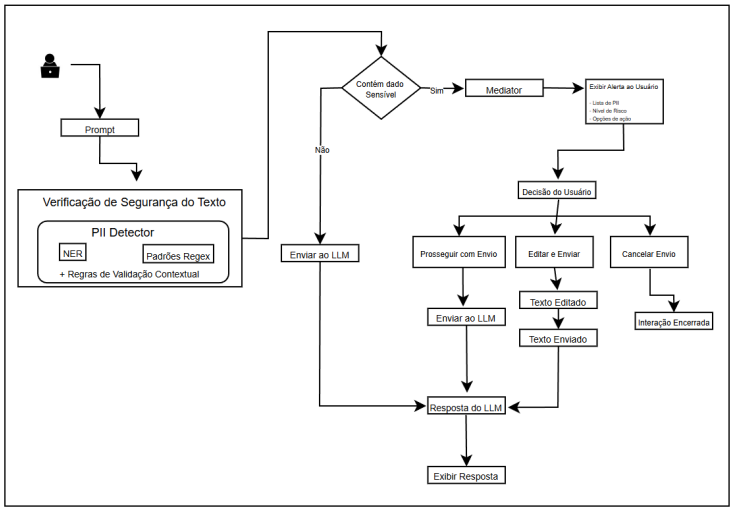

Figura 1. Componentes e relações da arquitetura da ferramenta de mediação de privacidade

A ferramenta proposta não realiza bloqueio automático da interação, mas atua como uma camada de mediação orientada à conscientização do usuário. Quando dados potencialmente sensíveis são identificados, o usuário pode revisar, editar, prosseguir ou cancelar o envio da mensagem (Angulo et al., 2011). Caso nenhuma ocorrência sensível seja detectada, a comunicação segue normalmente para o modelo de linguagem, preservando o fluxo original da aplicação.

3.2. Tecnologias de implementação

A ferramenta foi desenvolvida como uma extensão para o navegador Google Chrome, utilizando a plataforma de extensões do Chrome[1], TypeScript[2] e HTML[3], com suporte de *scripts* em *Shell* para automação do processo de desenvolvimento e *build*.

[1] `https://developer.chrome.com/docs/extensions/`
[2] `https://www.typescriptlang.org/`
[3] `https://developer.mozilla.org/en-US/docs/Web/HTML`

A ferramenta atua sobre interfaces web conversacionais, utilizando APIs da plataforma Chrome para interceptar campos de entrada, monitorar eventos de envio e inserir elementos de mediação. Essa estratégia incorpora a mediação de privacidade ao fluxo de uso sem exigir alterações estruturais. (Angulo et al., 2011).

O TypeScript foi adotado como principal linguagem de desenvolvimento devido ao suporte à tipagem estática, à integração com o ecossistema web e à facilidade de manutenção do código. Sua utilização ocorreu na implementação da lógica da ferramenta, incluindo o processamento do texto inserido, a aplicação das regras de detecção e o acionamento dos mecanismos de mediação em tempo real (Microsoft, 2024). O HTML foi empregado na estruturação dos componentes visuais, enquanto as bibliotecas *compromise.js*[4] e *compromise-dates*[5] apoiaram o processamento de linguagem natural e a identificação de entidades nomeadas e referências temporais. Além disso, foram utilizadas regras regex para padrões estruturados, como CPF, e-mail e números de telefone, e recursos nativos do *Document Object Model* (DOM)[6], empregados no acesso, inspeção e modificação dinâmica da página web (Elsevier, 2026).

A identificação de informações pessoais no texto inserido é realizada por um módulo de detecção que combina regras baseadas em expressões regulares e técnicas de reconhecimento de entidades nomeadas (*Named Entity Recognition* – NER). Os resultados obtidos são então encaminhados a um componente de mediação responsável por apresentar alertas *just-in-time* ao usuário (Mina et al., 2024).

3.3. Interface e interação da aplicação

A interação com a ferramenta começa na interface do *chatbot*, que não é um agente externo que se comunica com a aplicação proposta. A Figura 2 apresenta um exemplo da interface do ChatGPT. O usuário insere o texto normalmente no campo de entrada da aplicação, sem qualquer intervenção inicial por parte da ferramenta proposta, até o momento de envio da mensagem.

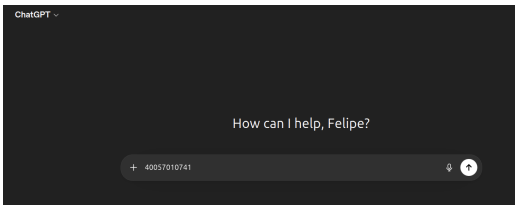

Figura 2. Inserção de um número de CPF na interface do ChatGPT antes do envio da mensagem ao modelo.

Em caso positivo, o envio é interrompido e uma interface de mediação é apresentada ao usuário, conforme ilustrado na Figura 3. Nessa etapa, a interface informa ao usuário o tipo de dado detectado, o trecho correspondente e o nível de risco atribuído à ocorrência. Essa classificação foi definida de forma heurística, considerando a natu-

[4] https://github.com/spencermountain/compromise

[5] https://github.com/spencermountain/compromise/tree/master/plugins/dates

[6] https://developer.mozilla.org/en-US/docs/Web/API/Document_Object_Model

reza do dado identificado, seu potencial de identificação direta do titular e os possíveis impactos de sua exposição indevida.

Assim, dados com maior potencial de fraude, uso indevido de identidade ou prejuízo financeiro, como CPF, RG e número de cartão de crédito, tendem a ser classificados em níveis mais elevados, enquanto dados que, quando considerados isoladamente, apresentam menor potencial de dano, podem receber classificações intermediárias ou baixas. Além da sinalização visual do risco, o alerta apresenta:

- Explicação resumida sobre a relevância da ocorrência para privacidade e proteção de dados pessoais;
- *Link* para informações adicionais sobre o tipo de dado detectado;

Esses elementos informacionais foram definidos para apresentar o risco identificado de forma clara e contextualizada, sem interromper automaticamente o fluxo de uso. A explicação resumida indica a relevância da ocorrência para a privacidade e a proteção de dados pessoais, enquanto os *links* permitem aprofundar informações sobre cada categoria de dado sem sobrecarregar a interface principal (Betzing et al., 2019).

A interface também apresenta o conteúdo digitado e disponibiliza uma área para edição antes do envio. Esse mecanismo foi projetado para apoiar a reflexão e a tomada de decisão informada, sem bloquear a continuidade da interação. Dessa forma, a aplicação preserva a autonomia do usuário e introduz uma camada adicional de proteção preventiva.

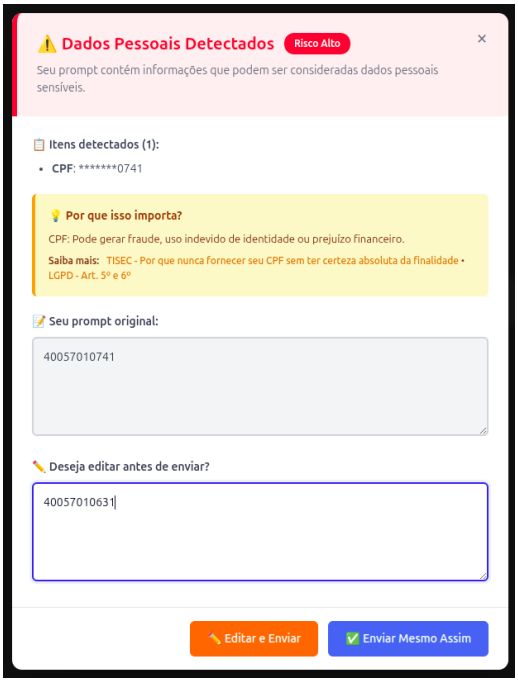

Figura 3. Interface de mediação após detecção de dado sensível.

Após a substituição da informação, o conteúdo revisado é apresentado no campo de texto do *chatbot*, conforme ilustrado na Figura 4. Esse comportamento evidencia a finalidade da proposta: estimular práticas mais conscientes de compartilhamento de informação sem interromper o fluxo da aplicação. Após a revisão, a interação pode prosseguir normalmente (Figura 5).

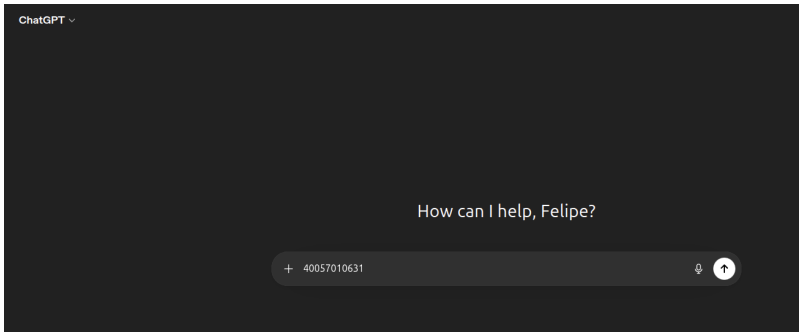

Figura 4. Conteúdo revisado após a mediação de privacidade.

Figura 5. Conteúdo enviado ao modelo após a revisão

4. Considerações Finais

Este artigo apresentou a proposta de uma ferramenta desenvolvida para identificar dados pessoais e sensíveis no conteúdo digitado em uma ferramenta conversacional e apresentar alertas visuais *just-in-time*, oferecendo ao usuário a possibilidade de revisar, editar ou manter a mensagem antes de seu envio. Diferentemente de abordagens baseadas em políticas extensas ou mecanismos pouco perceptíveis, a proposta torna o risco visível no momento da interação, favorecendo uma experiência mais consciente, transparente e alinhada a princípios de proteção de dados.

Além de ampliar a percepção sobre dados potencialmente sensíveis nas mensagens digitadas, a solução preserva a autonomia do usuário, evitando bloqueios automáticos e mantendo a decisão final sob seu controle. Com isso, o artefato apresentado pode contribuir para o desenvolvimento de aplicações conversacionais mais responsáveis, especialmente em contextos nos quais a interação com LLMs pode favorecer o compartilhamento excessivo ou inadvertido de dados pessoais.

Como limitação, destaca-se que a ferramenta ainda não foi avaliada com usuários em cenários reais de uso, o que impede afirmar sua eficácia na redução do compartilhamento de dados pessoais. Além disso, a abordagem baseada em expressões regulares e reconhecimento de entidades nomeadas pode apresentar falsos positivos e falsos negativos, especialmente em situações de ambiguidade semântica, variações linguísticas ou informações sensíveis dependentes de contexto. A solução também depende da estrutura DOM das interfaces analisadas, podendo exigir adaptações para diferentes plataformas.

Durante o desenvolvimento, observou-se que mecanismos de mediação precisam equilibrar proteção e fluidez da interação. Alertas excessivos podem comprometer a

experiência do usuário, enquanto alertas insuficientes podem reduzir a efetividade da proteção. Além disso, padrões estruturados, como CPF e e-mail, mostraram-se mais adequados para detecção por regras, enquanto dados contextuais exigem técnicas complementares de processamento de linguagem natural.

Embora os resultados preliminares sejam promissores, estudos futuros pretendem investigar a eficácia da solução na redução da exposição de dados pessoais e seus impactos sobre experiência, confiança e percepção de controle do usuário.

Como continuidade, pretende-se ampliar os dados detectados e a compatibilidade com diferentes plataformas conversacionais.

5. Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná pelo apoio financeiro concedido a esta pesquisa.

6. Declaração

Os autores declaram que utilizaram Inteligência Artificial Generativa para revisão textual (ortografia e gramática). Todo o conteúdo do texto foi gerado pelos autores.

Referências

Angulo, J., Fischer-Hübner, S., Pulls, T., and Wästlund, E. (2011). Towards usable privacy policy display & management: The primelife approach. In *Privacy and Identity Management for Life*, pages 108–118. Springer.

Baroni, L. and Pereira, R. (2024). Deceptive patterns under a sociotechnical view. In *Anais do XXIII Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais*, pages 459–471, Porto Alegre, RS, Brasil. SBC.

Betzing, J. H., Tietz, M., vom Brocke, J., and Becker, J. (2019). The impact of transparency on mobile privacy decision making. *Electronic Markets*, 30(3):607–625.

Chen, C., Zhou, D., Ye, Y., Li, T. J.-J., and Yao, Y. (2025). Clear: Towards contextual llm-empowered privacy policy analysis and risk generation for large language model applications. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25', page 277–297. Association for Computing Machinery.

Elsevier (2026). Document object model. ScienceDirect Topics. Acesso em: 28 mar. 2026.

European Data Protection Supervisor (2023). Generative ai: The data protection implications. Technical report, European Data Protection Supervisor (EDPS).

Freiberger, V., Fleig, A., and Buchmann, E. (2025a). "you don't need a university degree to comprehend data protection this way": Llm-powered interactive privacy policy assessment. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. ACM.

Freiberger, V., Fleig, A., and Buchmann, E. (2025b). "you don't need a university degree to comprehend data protection this way": Llm-powered interactive privacy policy assessment. In *CHI EA '25: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–12. Association for Computing Machinery. Published: 25 April 2025.

Ischen, C., Araujo, T., Voorveld, H., van Noort, G., and Smit, E. (2020). Privacy concerns in chatbot interactions. In *Chatbot Research and Design*, volume 11970 of *Lecture Notes in Computer Science*, pages 34–48. Springer.

Keraghel, I. and Nadif, M. (2025). Named entity recognition in the era of large language models: A comparative study. In *2025 International Conference on Advanced Machine Learning and Data Science (AMLDS)*, pages 617–623.

Leschanowsky, A., Rech, S., Popp, B., and Bäckström, T. (2024). Evaluating privacy, security, and trust perceptions in conversational ai: A systematic review. *Computers in Human Behavior*, 159:108344.

Li, T., Das, S., Lee, H.-P. H., Wang, D., Yao, B., and Zhang, Z. (2024). Human-centered privacy research in the age of large language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Lima, D., Vaz, N., Carvalho, S., and Berretta, L. (2025). Usabilidade de interfaces conversacionais com inteligência artificial generativa em aplicações mhealth: Uma revisão sistemática. In *Anais da XIII Escola Regional de Informática de Goiás*, pages 129–138, Porto Alegre, RS, Brasil. SBC.

Microsoft (2024). Typescript. TypeScript is a strongly typed programming language that builds on JavaScript, giving you better tooling at any scale.

Mina, M., Rodriguez-Penagos, C., Gonzalez-Agirre, A., and Villegas, M. (2024). Extending off-the-shelf ner systems to personal information detection in dialogues with a virtual agent: Findings from a real-life use case. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 44–53.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2025). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):106:1–106:72.

Seeger, A.-M., Pfeiffer, J., and Heinzl, A. (2021). Texting with human-like conversational agents: Designing for anthropomorphism. *Journal of the Association for Information Systems*, 22(4):1–58.

Yadav, V. and Bethard, S. (2018). A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zhang, Z., Jia, M., Lee, H.-P., Yao, B., Das, S., Lerner, A., Wang, D., and Li, T. (2024). "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery.

Zhou, J., Xu, E., Wu, Y., and Li, T. (2025). Rescriber: Smaller-llm-powered user-led data minimization for llm-based chatbots. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery.