Middleware para Interceptação de Dados Sensíveis em Chatbots com Modelos de Linguagem de Grande Porte (LLMs)

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 propositon of a technical solution for issuing preventive alerts during the display of responses generated by chatbots based on Large Language Models (LLMs). The technique consists of intercepting the content generated by the model at the time of display, with the aim of identifying the presence of potentially sensitive terms, in accordance with the principles of the LGPD, and displaying visual alerts to the user.*

Resumo. *Este artigo apresenta a proposta de solução técnica para a emissão de alertas preventivos durante a exibição de respostas geradas por chatbots baseados em Modelos de Linguagem de Grande Porte (LLMs). A técnica consiste na interceptação do conteúdo gerado pelo modelo no momento da exibição, com o objetivo de identificar a presença de termos potencialmente sensíveis, conforme os princípios da LGPD, e exibir alertas visuais ao usuário.*

1. Introdução

Os Modelos de Linguagem de Grande Porte (LLMs) são modelos computacionais baseados em inteligência artificial projetados para compreender e gerar texto semelhante ao humano, apresentando desempenho notável em diversas tarefas de processamento de linguagem natural (Raza et al., 2025). Modelos, como o GPT-4 (OpenAI), Claude (Anthropic), Command R+ (Cohere), Mixtral (Mistral) e Gemini (Google DeepMind), utilizam arquiteturas baseadas em mecanismos de atenção e são pré-treinados em grandes volumes de dados não rotulados, o que lhes confere notável fluência, adaptabilidade e capacidade de generalização contextual (Minaee et al., 2025).

Segundo Li et al. (2024), essa comunicação entre usuário e interfaces conversacionais baseados em LLMs tem trazido preocupações relacionadas à exposição de dados sensíveis durante as interações. Termos pessoais ou identificáveis podem ser processados e exibidos sem qualquer tipo de controle ou alerta contextual. As abordagens tradicionais de proteção, como políticas de privacidade ou consentimento inicial, mostram-se limitadas para lidar com essa exposição dinâmica, uma vez que não oferecem mecanismos de alerta no momento da interação.

Com base nesse cenário, este trabalho apresenta a propota de uma técnica para emissão de alertas preventivos durante a exibição de respostas geradas por chatbots baseados em LLMs. A técnica proposta consiste na interceptação do conteúdo gerado pelo modelo antes de sua apresentação ao usuário, com o objetivo de analisar a resposta quanto à presença de termos potencialmente sensíveis e exibir alertas visuais fundamentados nos princípios da LGPD. Esses alertas visam aumentar a transparência e promover a conscientização sobre riscos à privacidade, sem comprometer a continuidade da interação conversacional.

2. Fundamentação Teórica

Esta seção apresenta conceitos básicos que amparam esta pesquisa. Também são apresentados trabalhos relacionados.

2.1. *Transformers* e Modelos de Linguagem de Grande Porte (LLMs)

A evolução do Aprendizado de Máquina, do Processamento de Linguagem Natural (PLN) e de outras tecnologias de inteligência artificial, permitiu que os chatbots evoluíssem de sistemas baseados em regras fixas para modelos mais sofisticados. Destaca-se o uso dos *Transformers*, uma arquitetura que permite ao modelo analisar toda a sequência de palavras ao mesmo tempo e identificar quais partes são mais relevantes para entender o contexto. Isso tornou possível criar agentes conversacionais mais precisos, capazes de aprender com poucos exemplos (*few-shot*) ou até sem nenhum exemplo específico (*zero-shot*), alcançando alto desempenho em diversas tarefas (Patwardhan et al., 2023).

A adoção da arquitetura *Transformer* serve como base para os Modelos de Linguagem de Grande Porte (LLMs), que são treinados com extensos volumes de dados textuais e projetados para compreender, gerar e traduzir linguagem humana. Esses modelos têm se tornado centrais no desenvolvimento de sistemas conversacionais adaptativos, como o GPT, que utiliza camadas empilhadas de mecanismos de autoatenção e é pré-treinado em grandes conjuntos de dados não rotulados (Jana et al., 2024).

2.2. Exposição de Dados Sensíveis em Sistemas com LLMs

Respostas geradas por LLMs podem inadvertidamente revelar informações privadas durante a fase de inferência, mesmo sem que esses dados tenham sido explicitamente solicitados ou coletados. De acordo com Barberá (2025), os riscos à privacidade se manifestam em diferentes etapas do ciclo de vida dos sistemas baseados em LLMs, em especial na coleta e preparação de dados de treinamento, que podem incluir informações pessoais identificáveis ou sensíveis, e no monitoramento operacional, cujos logs de interação podem armazenar conteúdos que favorecem o vazamento ou uso indevido de dados.

Os riscos à privacidade se estendem também à fase de execução, em que, mesmo quando regras explícitas de proteção são configuradas, arquivos sensíveis podem ser expostos por meio de comandos maliciosos inseridos nos prompts ou pela interceptação do tráfego de rede. Em um número significativo de casos, conteúdos confidenciais, como relatórios técnicos, registros financeiros e documentos internos, foram recuperados a partir das respostas geradas ou dos pacotes transmitidos na sessão. Esses vazamentos decorrem de falhas nos mecanismos de controle implementados na lógica de acesso aos arquivos e evidenciam que a exposição de dados sensíveis pode ocorrer mesmo quando não há coleta direta ou armazenamento intencional dessas informações pelo modelo (Yan et al., 2025).

Segundo Greshake et al. (2023), os riscos à privacidade também podem ocorrer na fase de execução, quando, mesmo com regras explícitas de proteção, arquivos sensíveis são expostos por meio de comandos maliciosos inseridos nos *prompts*, um caso de engenharia de *prompt* indireta em que instruções inseridas em dados recuperados na inferência podem manipular o modelo e alterar seu comportamento.

2.3. Transparência na Interação com Modelos de Linguagem de Grande Porte em Chatbots

À medida que os LLMs influenciam tarefas cotidianas, muitos usuários não percebem que suas informações pessoais estão sendo coletadas e processadas durante a interação (Li et al., 2024). Esse desalinhamento entre o tratamento de dados e a percepção do usuário compromete o princípio da transparência previsto na LGPD, segundo o qual o titular deve ser informado de forma clara, acessível e contínua sobre o uso de seus dados pessoais (LGPD - Brasil, 2018).

Contudo, a própria estrutura desses sistemas dificulta a oferta de mecanismos informativos durante a interação. Interfaces conversacionais baseadas em LLMs raramente comunicam em tempo real quais dados estão sendo inferidos ou utilizados, o que favorece a persistência de assimetrias informacionais (Łajewska et al., 2024).

2.4. Trabalhos Relacionados

Para elaborar o mecanismo de alerta proposto, tornou-se necessário realizar um estudo preliminar exploratório sobre abordagens de transparência e mitigação de riscos em chatbots baseados em LLMs, buscando identificar quais métodos e diretrizes têm sido empregados para detectar dados sensíveis e apresentar avisos oportunos ao usuário durante a interação.

Entre os estudos analisados, Mireshghallah et al. (2024) demonstraram que diálogos públicos com modelos GPT apresentam incidência expressiva de informações sensíveis. Nesse estudo, as ameaças foram classificadas em memorização, inferência e divulgação, revelando que detalhes de saúde, preferências pessoais e informações financeiras emergem mesmo sem solicitação explícita. Constatou-se, ainda, que os usuários raramente percebem esses vazamentos, o que evidencia a limitação de abordagens centradas unicamente na filtragem de Informações Pessoais Identificáveis (PII) e reforça a necessidade de alertas contextuais durante a conversação.

Em outra linha de investigação, Freiberger et al. (2025) apresentaram a extensão PRISMe (Privacy Risk Information Scanner for Me), um add-on para navegador que usa um LLM para inspecionar, durante a navegação, as políticas de privacidade das páginas visitadas, sinalizando riscos por meio de ícones coloridos e oferecendo resumos interativos em um painel e um chat. Em estudo qualitativo com 22 participantes, o uso da ferramenta elevou a consciência sobre privacidade e facilitou a compreensão dos termos, embora o protótipo ainda se restrinja a documentos estáticos, sem analisar o conteúdo gerado em tempo real nas conversas com chatbots.

Em suma, a literatura aponta que há riscos de exposição de dados em interações com LLMs e que ferramentas de explicação em tempo real podem aumentar a conscientização dos usuários. Adicionalmente, há abordagens que podem contribuir na

integração e detecção automática de dados sensíveis dentro do fluxo conversacional a um mecanismo de alerta direcionado ao usuário.

3. Metodologia

Na Figura 1 é apresentado o fluxograma da técnica proposta, que contempla as seguintes ações: (1) envio do prompt pelo usuário; (2) geração da resposta pelo modelo de linguagem; (3) interceptação da resposta antes da exibição; (4) verificação da presença de dados sensíveis; (5) decisão sobre exibir alerta; e (6) continuidade ou encerramento da interação conforme a escolha do usuário.

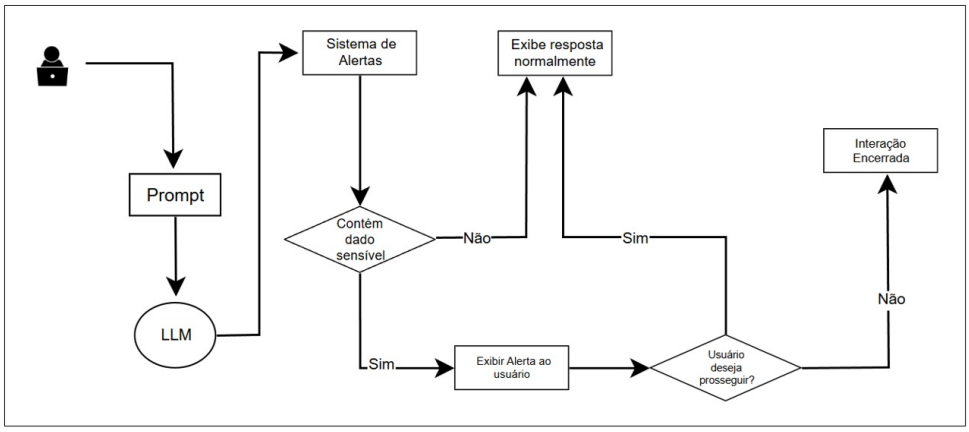

Figura 1. Fluxo da técnica de alerta preventivo em sistemas conversacionais baseados em LLMs.

Pretende-se implementar a proposta como um componente intermediário (*middleware*), acoplado à interface do sistema conversacional. Ao receber a resposta do LLM, esse componente realiza a interceptação do conteúdo e o submete a um mecanismo de verificação de sensibilidade, cuja lógica baseia-se em listas de termos e expressões extraídas de um léxico de dados sensíveis, construído com base nas categorias previstas no art. 5º da LGPD.

Caso a resposta contenha indícios de dados sensíveis, o sistema aciona um módulo de exibição de alerta visual, informando o risco ao usuário em tempo real. O alerta é exibido de forma não intrusiva, e a decisão de seguir ou não com a interação permanece sob controle do usuário. Se não forem encontrados dados sensíveis, a resposta é exibida normalmente.

Após a implementação da técnica por meio do sistema, o próximo passo será a avaliação em ambiente simulado, com base em cenários diversos, para verificar a efetividade da sinalização e o impacto sobre a experiência de uso.

4. Conclusão

Este trabalho apresentou uma proposta de técnica para alerta preventivo durante a exibição de respostas geradas por chatbots baseados em LLMs. A abordagem visa mitigar riscos à privacidade por meio da identificação de termos potencialmente sensíveis antes da entrega da resposta ao usuário, atendendo aos princípios de transparência e prevenção previstos na LGPD.

A técnica consiste na interceptação do conteúdo gerado e sua análise lexical com base em categorias legais, acionando um alerta visual sempre que identificado risco de exposição indevida. O objetivo é ampliar a consciência do usuário sobre o tratamento de suas informações, preservando a continuidade da interação.

O presente trabalho configura-se como uma proposta conceitual e encontra-se em estágio inicial de desenvolvimento, não tendo sido ainda submetido a avaliação empírica. A próxima etapa compreenderá a implementação da solução e a realização de testes controlados em ambiente simulado para mensurar sua acurácia, as taxas de falsos positivos e de falsos negativos, bem como o impacto percebido na experiência de uso.

Dessa forma, espera-se que a solução contribua para o debate sobre transparência e proteção de dados em chats baseados em LLMs. Como próximos passos, propõe-se a realização de testes em ambiente simulado com diferentes cenários de uso, a fim de verificar a efetividade da sinalização e o impacto percebido sobre a experiência do usuário.

Referências

Barberá, I. (2025). AI Privacy Risks & Mitigations – Large Language Models (LLMs). `https://edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf`. Commissioned by the European Data Protection Board (EDPB) under the Support Pool of Experts (SPE). Views expressed are those of the author.

Freiberger, V., Fleig, A., and Buchmann, E. (2025). "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.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., and Fritz, M. (2023). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, page 79–90, New York, NY, USA. Association for Computing Machinery.

Jana, S., Biswas, R., Pal, K., Biswas, S., and Roy, K. (2024). The evolution and impact of large language model systems: A comprehensive analysis. *Alochana Journal*, 13(3):65–78.

LGPD - Brasil (2018). Lei nº 13.709, de 14 de agosto de 2018: Lei Geral de Proteção de Dados Pessoais (LGPD). `http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`. Diário Oficial da União, Seção 1, 15 ago. 2018.

Li, T., Das, S., Lee, H.-P., 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)*, pages 1–4. Association for Computing Machinery.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2025). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Mireshghallah, N., Antoniak, M., More, Y., Choi, Y., and Farnadi, G. (2024). Trust no bot: Discovering personal disclosures in human-llm conversations in the wild. *arXiv*

preprint arXiv:2407.11438. Accepted at COLM 2024. Version 2, last revised 20 Jul 2024.

Patwardhan, N., Marrone, S., and Sansone, C. (2023). Transformers in the real world: A survey on nlp applications. *Information*, 14(4).

Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., and Sattar, M. A. (2025). Industrial applications of large language models. *Scientific Reports*, 15(1):13755.

Yan, C., Guan, B., Li, Y., Meng, M. H., Wan, L., and Bai, G. (2025). Understanding and detecting file knowledge leakage in gpt app ecosystem. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 3831–3839, New York, NY, USA. Association for Computing Machinery.

Łajewska, W., Spina, D., Trippas, J., and Balog, K. (2024). Explainability for transparent conversational information-seeking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 1040–1050. ACM.