Estado da Arte II e Metodologia: Segurança, Privacidade e Conformidade em Aplicações com LLMs

Leonardo Nunes[1], **Antônio Marcos**[1], **Álvaro Gueiros**[1], **Lucas William**[1], **Mauro Vinícius**[1], **Vandielson Tenório**[1]

[1]Aluno da disciplina de Segurança da Informação do Bacharelado em Ciência da Computação – Universidade Federal do Agreste de Pernambuco (UFAPE)

Abstract. *This paper advances the second step of a literature review and formalizes the methodology for a proof-of-concept on LLM security. From three complementary sources—risk and privacy challenges, adaptive access control in healthcare, and AI Act compliance guidance—we identify a gap: the lack of an end-to-end, reproducible evaluation framework that jointly measures technical mitigations (e.g., LLM firewall, RAG, sanitization), adaptive RBAC, and evidence of regulatory compliance. We propose an architecture and experimental plan (A/B and ablation) with multi-metric assessment (attack success, precision/recall/F1, latency, cost, and compliance coverage) to fill this gap.*

Resumo. *Este artigo consolida a segunda etapa da revisão bibliográfica e formaliza a metodologia para um proof-of-concept em segurança de LLMs. A partir de três fontes complementares—desafios de privacidade e segurança, controle de acesso adaptativo em saúde e diretrizes de conformidade ao AI Act—identificamos a lacuna: ausência de um framework avaliativo end-to-end, reprodutível, que integre mitigações técnicas (firewall LLM, RAG, sanitização), RBAC adaptativo e evidências de conformidade. Propomos arquitetura e desenho experimental (A/B e ablação) com avaliação multi-métrica (taxa de sucesso de ataque, precisão/recall/F1, latência, custo e cobertura de compliance) para preencher essa lacuna.*

1. Introdução

Modelos de linguagem de grande porte (LLMs) ampliaram capacidades de automação e suporte à decisão, mas introduziram novas superfícies de ataque (injeção e *indirect prompt injection*, *insecure output handling*, *denial-of-wallet*/DoS, vazamento de dados e vieses de saída) e responsabilidades regulatórias. A literatura recente oferece: (i) taxonomias de riscos e controles técnicos; (ii) evidências setoriais de controle de acesso adaptativo com ganhos mensuráveis; e (iii) traduções de requisitos regulatórios em ações implementáveis. Apesar disso, ainda falta uma avaliação integrada e padronizada que una esses três eixos em um mesmo experimento reprodutível.

Contribuições. (1) Identificação de uma lacuna de pesquisa end-to-end; (2) Proposta de arquitetura integrada (sanitização, firewall LLM, RAG, RBAC adaptativo, auditoria/mapeamento de conformidade); (3) Desenho experimental com testes A/B e ablação, e (4) conjunto de métricas para segurança, desempenho, custo e conformidade.

2. Lacuna de Pesquisa

Com base no levantamento de desafios de segurança e privacidade em LLMs (controles como *LLM firewall*, RAG, *differential privacy*, HITL) [1], no estudo de *RBAC* adaptativo com detecção de anomalias no domínio de saúde [2], e no guia prático de conformidade com o *EU AI Act* (papéis, controles e mapeamento a normas) [3], identificamos a seguinte lacuna:

Lacuna central: falta um **framework avaliativo end-to-end**, reprodutível e alinhado a normas, que combine em uma *mesma* aplicação de LLM: (i) mitigação técnica de riscos (injeção de *prompt*, *output handling*, DoS/*denial-of-wallet*), (ii) **controle de acesso adaptativo** (RBAC dinâmico com *risk score* e detecção de anomalias), (iii) **privacidade por design** (sanitização e RAG com repositório controlado), e (iv) **traçabilidade de conformidade** (AI Act/OWASP/ISO) com evidências objetivas. Hoje há sínteses conceituais, um caso setorial e diretrizes de compliance, porém *não* há avaliação comparativa padronizada do *conjunto* desses controles sob ataques realistas, com métricas unificadas de segurança, privacidade, custo e desempenho.

3. Trabalhos Relacionados

Levantamentos recentes sistematizam ameaças em LLMs (p. ex., *prompt injection*, vazamento, DoS, viés) e indicam controles como *LLM firewalls*, sanitização de entrada/saída, RAG, *differential privacy* e HITL [1]. Em paralelo, no domínio de saúde, [2] propõem *RBAC* adaptativo acoplado à detecção de anomalias assistida por LLM, com sanitização/redação de entidades sensíveis e avaliação quantitativa (acurácia, precisão, *recall*, F1) em dados sintéticos. No eixo regulatório, [3] traduzem o *EU AI Act* em ações implementáveis e mapeiam responsabilidades por papel (provider/hoster/integrator) e controles alinhados a OWASP/ISO/ENISA.

Síntese crítica. Esses trabalhos oferecem (i) taxonomia e controles, (ii) um caso setorial com ganhos medidos, e (iii) ponte normativa→ação. O passo ainda ausente é uma **avaliação integrada**—com *benchmark* reprodutível e métricas comparáveis—que una mitigação técnica, *RBAC* adaptativo e geração de evidências de conformidade em um *mesmo* pipeline experimental.

4. Tabela Comparativa dos Trabalhos

Tabela 1. Comparação dos trabalhos relacionados e evidência da lacuna

Eixo	**Rathod et al.** [1]	**Yarram et al.** [2]	**Bunzel** [3]	**Lacuna**
Ameaças mapeadas	Abrangente (injeção, vazamento, DoS, viés; princípios OWASP)	Foco em saúde; acessos e anomalias; avaliação empírica	Tradução AI Act → controles; papéis e responsabilidades	Integração prática + avaliação comparativa unificada
Controles	Firewall LLM, DP, RAG, HITL, sanitização E/S	RBAC dinâmico + detecção de anomalias; sanitização de *queries*	Playbook de compliance e matriz de riscos/controles	Arquitetura end-to-end com métricas padronizadas
Evidência experimental	Predominante conceitual/sintética	Resultados quantitativos vs. regras/assinaturas (A/P/R/F1)	Diretrizes sem *benchmark* técnico unificado	*Benchmark* reprodutível multi-métrica
Conformidade/regulação	Boas práticas e princípios	Menções a HIPAA/GDPR (alto nível)	Mapeia AI Act ↔ OWASP/ISO/ENISA	Evidências automáticas e rastreáveis de conformidade

5. Metodologia

5.1. Objetivo e Visão Geral

Projetar e avaliar um **pipeline** de segurança para um aplicativo com LLM (assistente de conhecimento institucional), integrando: **sanitização de entrada** → **LLM firewall** → **RAG** (base privada) → **RBAC adaptativo** (com *risk score*) → **sanitização de saída** → **auditoria & mapeamento de conformidade**.

5.2. Arquitetura do Protótipo (Fim-a-Fim)

1. **Sanitização de entrada**: NER/*redaction* de PII, *regex* e listas semânticas de bloqueio; normalização de formatos.
2. **LLM Firewall**: regras + detecção semântica de instruções adversariais; *deny-list* de capacidades perigosas; *rate-limiting*. ([1])
3. **RAG privado**: repositório controlado (documentos permitidos, metadados de confidencialidade/política, *caching* sob política) para reduzir memorizações e vazamento. ([1])
4. **RBAC adaptativo**: *risk score* por requisição (papel, horário, dispositivo, histórico, semântica da consulta); se risco $\geq$ limiar $\Rightarrow$ MFA/*step-up*/bloqueio. ([2])
5. **Sanitização de saída**: verificações de *policy*, encoding seguro e filtros de PII/código executável; **auditoria** *append-only*.
6. **Mapper de conformidade**: geração de evidências para artigos (p.ex., 9/10/11/15) do AI Act, alinhadas a OWASP/ISO/ENISA; explicita papéis e SLAs técnicos. ([3])

5.3. Ameaças e Cenários de Teste

- **Ameaças**: *prompt/indirect injection*, *insecure output handling*, *denial-of-wallet*/DoS, *model/knowledge stealing* por *querying*, *membership inference* (nível básico), abuso de papel e *break-glass*.
- **Desenho experimental**: **Testes A/B** e **ablação**:
    1. Baseline (sem controles);
    2. + Firewall LLM;
    3. + RAG privado;
    4. + RBAC adaptativo;
    5. **Pipeline completo** (todas as camadas).
- **Dados**: corpus institucional neutro + **dados sintéticos** no domínio de saúde (evitar PHI, variabilidade controlada), conforme práticas de [2].

5.4. Métricas e Coleta

- **Segurança/Privacidade**: *Attack Success Rate* (quanto menor, melhor), precisão, *recall* e F1 dos detectores; taxa de vazamento; eficácia de *rate-limiting*.
- **Desempenho/Custos**: latência p95/p99; custo por requisição; custo por bloqueio; *throughput* sob carga.
- **Conformidade**: % de requisitos cobertos (AI Act/OWASP/ISO), papéis definidos e logs exportáveis como evidência.

5.5. Critérios de Sucesso

Redução de $\geq X\%$ na taxa de sucesso de ataque com perda $\leq Y\%$ de qualidade; *overhead* de latência $\leq Z\%$; cobertura de compliance $\geq W\%$ com evidências auditáveis exportadas.

5.6. Reprodutibilidade e *Open Science*

Organização do repositório: `/src` (módulos de firewall, RAG, RBAC, sanitização, auditoria), `/eval` (ataques, cenários A/B, métricas), `/data` (amostras sintéticas e *policy store*), `/paper` (SBC LaTeX), `/slides` (Sprint Review). Scripts de execução e relatórios automatizados para reproduzir todos os experimentos.

6. Conclusão e Próximos Passos

Apresentamos a lacuna e um plano metodológico para avaliá-la de forma integrada, com métricas comparáveis e geração de evidências de conformidade. Como próximos passos: (i) implementação do protótipo, (ii) definição dos cenários de ataque e parâmetros de A/B, (iii) execução dos experimentos e análise dos *trade-offs*, e (iv) disponibilização pública dos artefatos e relatórios.

Referências

[1] Rathod, *et al.* (2024). Privacy and Security Challenges in Large Language Models.

[2] Yarram, *et al.* (2024). Privacy-Preserving Healthcare Data Security Using LLMs and Adaptive Access Control.

[3] Bunzel (2024). Compliance Made Practical: Translating the EU AI Act into Implementable Security Actions.