Estado da Arte II e Metodologia

Segurança, Privacidade e Conformidade em Aplicações com LLMs

Leonardo Nunes, Antônio Marcos, Álvaro Gueiros, Lucas William, Mauro Vinícius, Vandielson Tenório (UFAPE)

Objetivos da Apresentação

Consolidar Revisão

Síntese da literatura em segurança, privacidade e conformidade para LLMs

Lacuna Identificada

Framework avaliativo integrado, reprodutível e alinhado a normas regulatórias

Metodologia Proposta

Protótipo end-to-end com controles técnicos e evidências de compliance mensuráveis

Contexto e Motivação

Desafios Técnicos

- Novas superfícies de ataque (prompt injection, vazamento de dados)
- Negação de serviço (DoS/wallet depletion)
- Falta de controle granular sobre acessos

Pressões Regulatórias

- AI Act europeu e regulamentações globais emergentes
- Requisitos de auditoria e rastreabilidade
- Necessidade de métricas comparáveis e reprodutíveis

Estado da Arte — Fontes-Chave

1

Rathod et al.

Análise de riscos e controles técnicos: firewall semântico, RAG, sanitização, DP diferencial e intervenção humana (HITL)

2

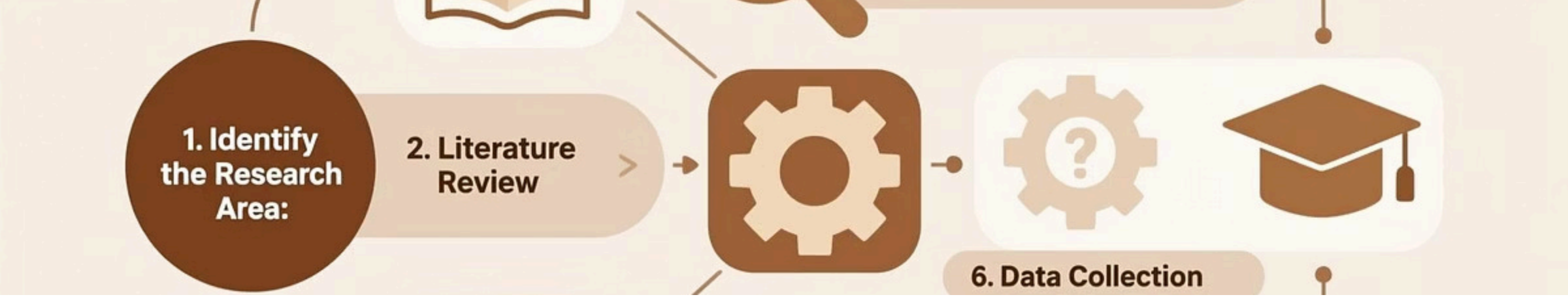
Yarram et al.

RBAC adaptativo e detecção de anomalias com validação no domínio da saúde e dados sensíveis

3

Bunzel

Mapeamento do AI Act para ações implementáveis: papéis, controles, auditorias e geração de evidências



Lacuna de Pesquisa

Falta um **framework avaliativo integrado, reprodutível e alinhado a normas**, que combine em uma única aplicação:

→ Mitigação Técnica

Firewall semântico, RAG privado, sanitização de entrada/saída

→ Controle de Acesso

RBAC adaptativo com scoring de risco por requisição

→ Conformidade Mensurável

Evidências quantitativas alinhadas a AI Act, OWASP e ISO

Hoje: sínteses e casos isolados, não avaliação integrada com benchmark multi-métrica.

Arquitetura do Protótipo — Pipeline End-to-End

Cada etapa aplica múltiplos controles em cascata, com rastreabilidade e logs para auditoria contínua.

Desenho Experimental

Cenários de Ataque

- Prompt injection direto e indireto
- Insecure output (código malicioso)
- Denial-of-wallet e abuso de papel
- Vazamento de informações sensíveis

Protocolo Experimental

Progressão:

1. Baseline (sem controles)
2. Apenas Firewall
3. Firewall + RAG
4. Completo com RBAC
5. Pipeline integrado + auditoria

Métricas e Critérios de Sucesso

↓X%

Attack Success Rate

Redução de ataques bloqueados com precisão e recall >90%

≤Y%

Degradação de Qualidade

Manutenção de respostas úteis com overhead aceitável

≤Z ms

Latência P95/P99

Throughput viável sem gargalos críticos

≥W%

Cobertura Compliance

Requisitos regulatórios rastreáveis e auditáveis

Valores X, Y, Z, W serão calibrados no piloto com base em critérios setoriais.

Entregáveis da Sprint

Diretório	Conteúdo
/src	Módulos: firewall, RAG, RBAC, sanitização, auditoria
/eval	Cenários de ataque, A/B testing, cálculo de métricas
/paper	SBC LaTeX: Trabalhos Relacionados + Metodologia
/slides	Apresentações e documentação visual (Gamma)

Status: rascunhos de Trabalhos Relacionados, Metodologia e tabela comparativa sincronizados.

Próximos Passos e Fechamento

1. Implementar MVP

Codificação do pipeline completo com integração de componentes

2. Experimentos A/B & Ablação

Validar impacto isolado de cada controle e combinações

3. Trade-offs & Análise

Segurança versus custo computacional e latência

4. Consolidar Evidências

Conformidade regulatória e potencial de generalização

Feedback solicitado: Validação das métricas, calibração de limiares e escopo de replicabilidade.