Segurança de Prompts em LLMs

Análise comparativa de abordagens de proteção

Panorama da Pesquisa

Cinco abordagens distintas para segurança em modelos de linguagem

Defesa Unificada

Framework integrado contra múltiplos ataques

Análise de Vulnerabilidades

Testes sistemáticos em 36 LLMs

Privacidade do Usuário

Proteção de dados em chatbots

Sistematização do conhecimento

Propondo uma taxonomia e um conjunto de métricas próprias para padronizar avaliações de ataques e defesas de LLMs.

UniGuardian: Defesa Unificada

Lin, Huawei et al.

Foco Principal

Framework de defesa contra:

- Injeção de prompts
- Ataques backdoor
- Ataques adversariais

Metodologia

Sistema unificado de detecção

Análise de prompts e saídas

Métricas do UniGuardian

auROC

Area Under the Receiver Operator Characteristic Curve

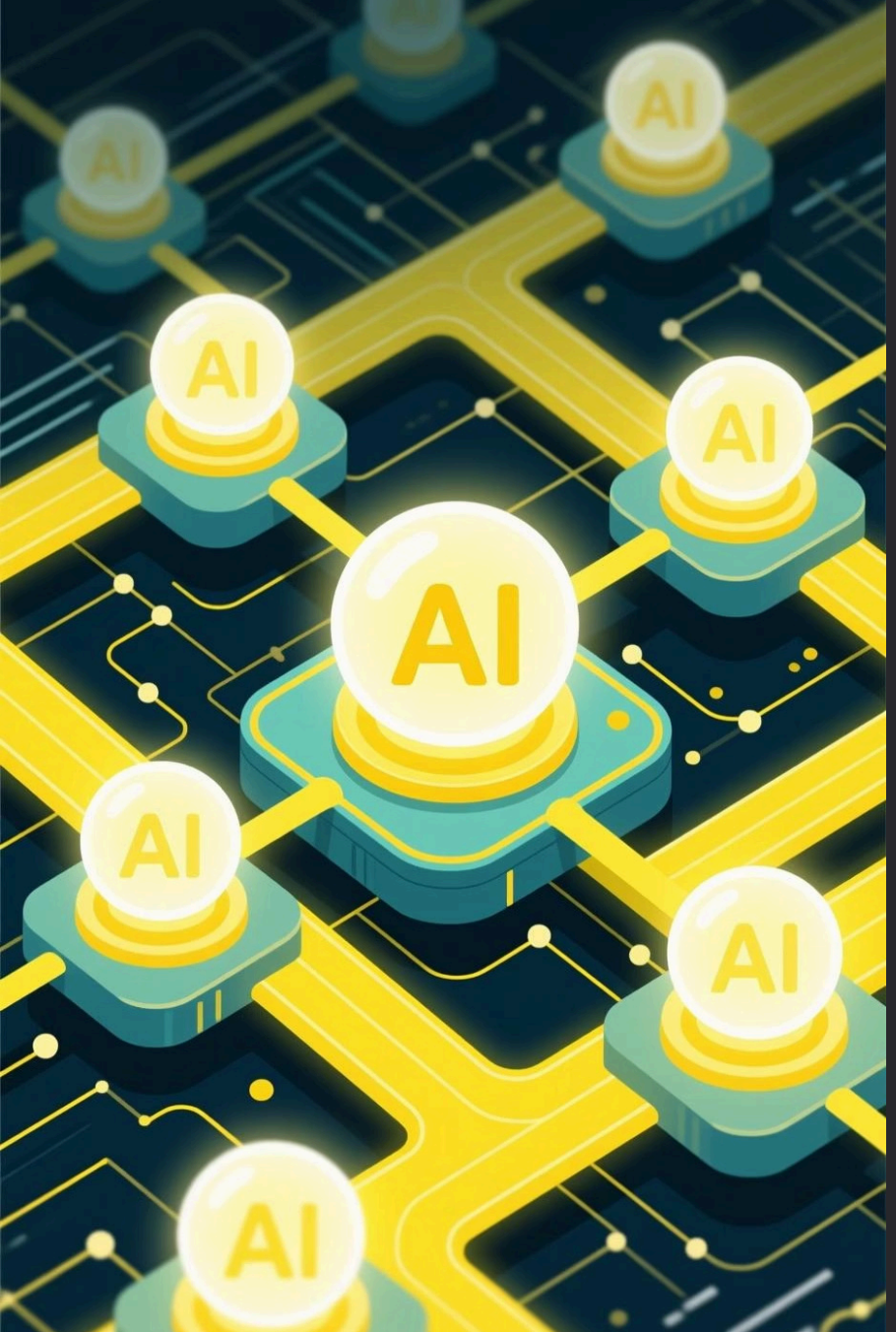
Mede capacidade de distinguir ataques

auPRC

Area Under the Precision-Recall Curve

Avalia precisão da detecção

Lacuna: Falsos positivos em ataques backdoor complexos ou ofuscados



Vulnerabilidades em Escala

Benjamin, Victoria et al.

36

LLMs Testados

Análise sistemática abrangente

4

Prompts de Injeção

Focados em keylogger

Metodologia de Análise

01

Injeção Direta

4 prompts contra 36 LLMs

02

Análise Estatística

Correlação e Random Forest

03

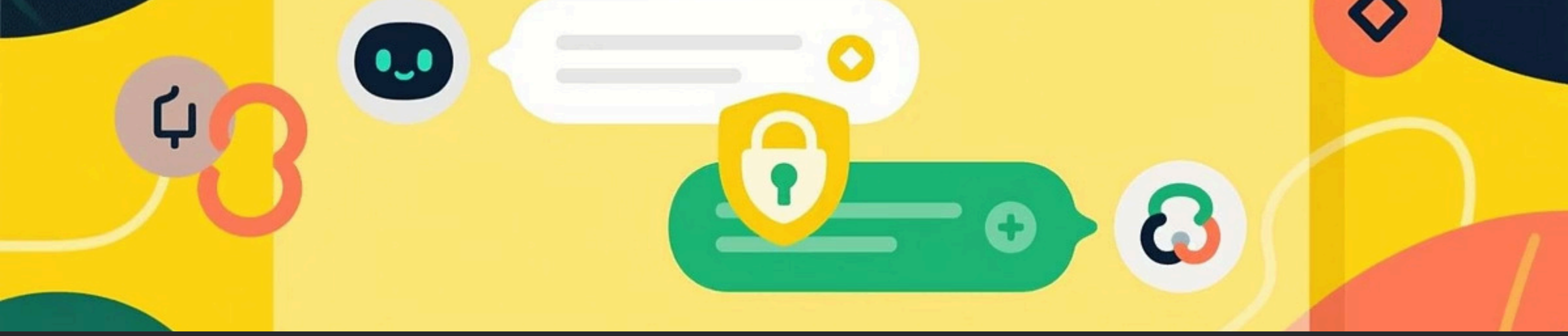
Interpretação

SHAP e PCA para features

Métricas Avaliadas

- Taxa de sucesso dos ataques
- Importância de features
- Correlação entre prompts

Lacuna: Necessidade de testes multilíngues e múltiplas etapas



Privacidade em Chatbots

Sebastian, Glorin

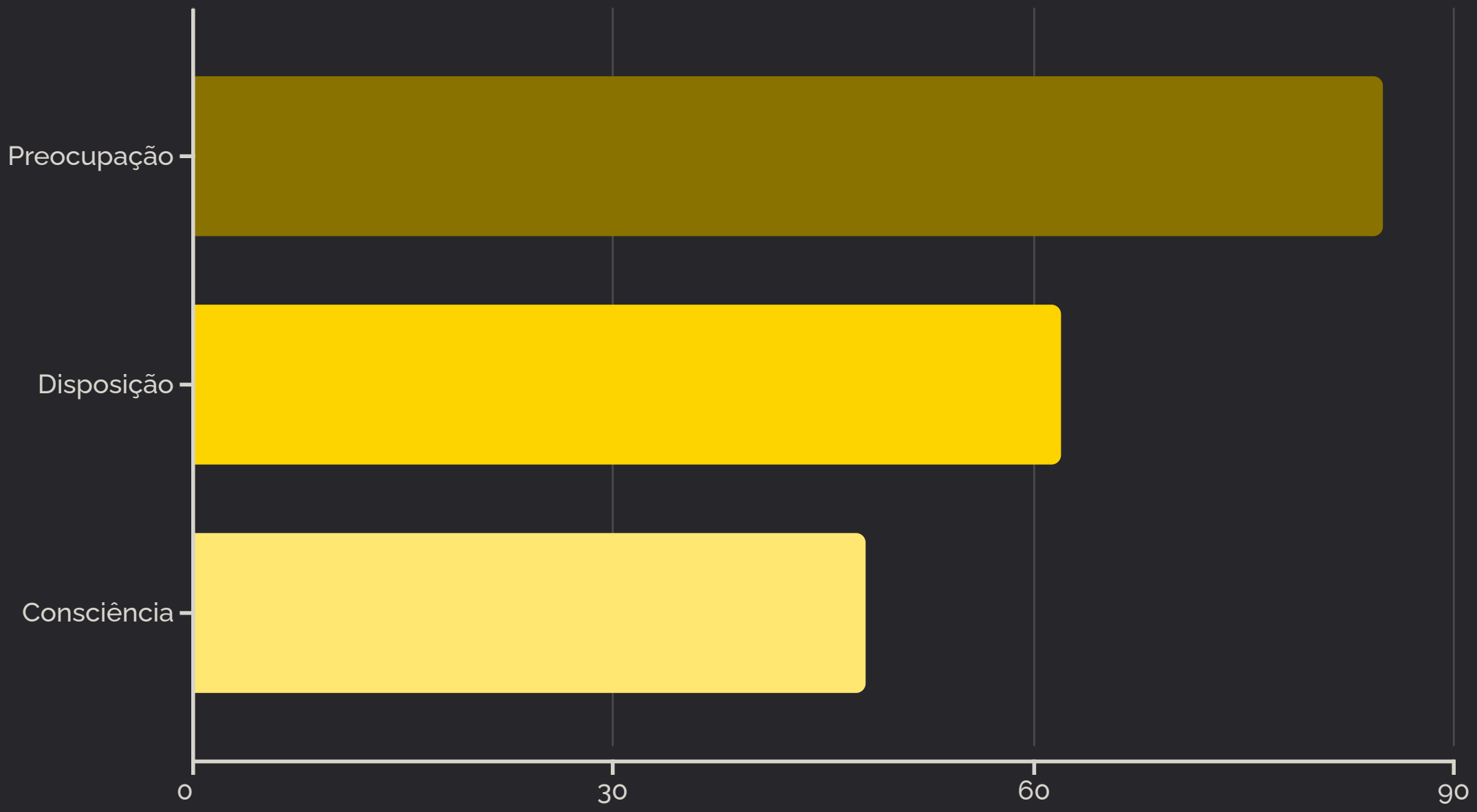
Investigação

Proteção de dados no ChatGPT

Pesquisa

177 usuários consultados

Percepção dos Usuários



Técnicas analisadas: privacidade diferencial e outras PETs

Lacuna: Análise de riscos à privacidade através de pesquisas com usuários

SoK Taxonomy and Evaluation of Prompt Security in Large Language Models

Hong, Hanbin

Foco Principal

Sistematização do conhecimento:

- Propondo 3 taxonomia (ataque, defesa e vulnerabilidade)
- Protocolos experimentais unificados baseados em um esquema (M, A, D, S e J).
- Grande banco de dados públicos de prompts rotulados (JailbreakDB).

Metodologia

Revisão sistemática e proposta de uma taxonomia multinível.

Métrica

Padronizadas próprias integradas a um toolkit de avaliação com taxonomia hierárquica e perfis de ameaça formais.

Lacuna: Limitação na aplicação prática das métricas e taxonomias propostas; necessidade de validação contínua e ampliação para modelos multimodais e cenários reais.

Comparativo de Abordagens

UniGuardian

Detecção unificada

auROC + auPRC

Benjamin et al.

Análise sistemática

Taxa de sucesso

Sebastian

Foco em privacidade

Percepção usuário

Hong et al.

Revisão Bibliográfica

Taxonomia, Métrica Própria

e Banco de dados

Lacunas e Oportunidades

Ataques Complexos

Backdoors ofuscados geram falsos positivos

Testes Multilíngues

Necessidade de validação em múltiplos idiomas

Próximos passos: Integração de abordagens para segurança holística