Atividade 6 - Prévia dos resultados

Tema: Segurança de Prompt em Modelos de LLM: Protegendo a Inteligência Conversacional contra Manipulações e Vazamentos

Alunos: Antonio Marcos, Álvaro Gueiros, Leonardo Nunes, Lucas William, Mauro Vinícius, Vandielson Tenório.

Resultados Parciais

Esta seção apresenta os resultados parciais dos experimentos realizados com o sistema de Chat Seguro para LLMs, avaliando quatro cenários:

- Prompt Seguro
- Prompt Injection 1
- Prompt Injection 2
- Prompt Longo Demais

As principais métricas de desempenho e segurança avaliadas incluíram:

- **Taxa de detecção:** Capacidade de identificar entradas maliciosas com precisão.
- Falsos positivos: Quantidade de entradas legítimas tratadas incorretamente como ataques.
- Latência média (ms): Tempo médio de processamento de cada requisição.
- Throughput (req/s): Número de requisições processadas por segundo, indicando a capacidade do sistema.

Análise Geral dos Resultados

Os experimentos confirmaram um desempenho **altamente satisfatório** do sistema de Chat Seguro para LLMs. A taxa de detecção alcançou **100% (1.0)** para prompts seguros e para simulações de ataques de **Prompt Injection**, demonstrando a eficácia do mecanismo de segurança em identificar e neutralizar a manipulação maliciosa do modelo.

Um ponto que merece atenção é a ocorrência de **apenas um falso positivo**, registrado no cenário de Prompt Seguro. Este incidente sinaliza a necessidade de **ajustes finos nos critérios de classificação** para reduzir bloqueios indevidos de entradas legítimas, sem comprometer a segurança geral da plataforma.

Discussão dos Cenários Testados: Prompt Seguro

No cenário de **Prompt Seguro**, foi registrado apenas um falso positivo. Este incidente indica uma leve sensibilidade excessiva do sistema, atribuída a heurísticas de segurança conservadoras.

Desempenho Técnico

- **Latência:** 13,28 ms
- **Throughput:** 75,28 req/s

O sistema opera de forma eficiente, sem comprometer a experiência do usuário com atrasos significativos.

Recomendação

O sistema é funcional e seguro, mas um refinamento adicional é recomendado para otimizar a precisão. O objetivo é evitar classificações incorretas de prompts legítimos, balanceando a detecção de ameaças com a minimização de falsos positivos e aprimorando a experiência do usuário sem sacrificar a segurança.

Discussão dos Cenários Testados: Prompt Injection 1

No cenário de **Prompt Injection 1**, o sistema demonstrou eficácia total, **detectando integralmente o ataque (100% de taxa de detecção)**, e sem registrar **nenhum falso positivo**. Isso valida a precisão do mecanismo de segurança.

Em termos de desempenho, a latência média foi **reduzida para 8,73 ms** e o throughput **aumentou para 114,6 req/s**. Estes números indicam que a detecção de ataques é computacionalmente leve e permite que o sistema mantenha alta performance mesmo sob tentativas de ataque.

Em resumo, o algoritmo de detecção é **altamente eficiente, seguro e otimizado**, protegendo os LLMs contra manipulações sem comprometer o desempenho.

Discussão dos Cenários Testados: Prompt Injection 2

No cenário de **Prompt Injection 2**, o sistema reconfirmou sua **robustez e consistência**. A detecção de ataques foi **perfeita (100%)**, com ausência total de falsos positivos, sublinhando a eficácia do mecanismo defensivo contra variações de ataques de Prompt Injection.

Em termos de desempenho, a latência foi ainda menor, registrando **8,56 ms**, e o throughput aumentou para **116,76 req/s**. Esses números demonstram não só a manutenção da alta performance, mas também uma otimização contínua em relação ao cenário anterior.

Os resultados reiteram a **confiabilidade e a robustez** do sistema de segurança, protegendo consistentemente os LLMs contra diferentes tipos de Prompt Injection e consolidando sua eficácia como uma barreira defensiva.

Discussão dos Cenários Testados: Prompt Longo Demais

O teste de **Prompt Longo Demais** avaliou a capacidade do sistema de lidar com entradas que excedem os limites técnicos, prevenindo potenciais ataques de negação de serviço. O sistema demonstrou eficácia total, identificando corretamente todos os prompts inválidos com **100% de detecção e zero falsos positivos**.

O desempenho foi excelente, com uma latência de **8,97 ms** e um throughput de **111,46 req/s**. Isso comprova que o bloqueio de entradas excessivamente longas é altamente eficiente e não afeta a performance geral do sistema.

Este resultado é crucial, pois confirma que o sistema é **extremamente eficaz** em proteger os LLMs contra prompts abusivos ou que extrapolam limites técnicos, garantindo sua estabilidade e disponibilidade, mesmo sob tentativas de sobrecarga.

Considerações sobre Desempenho

A análise da latência e do throughput em todos os cenários revela um desempenho **extremamente rápido e eficiente**.

- **Média de Latência:** Entre 8 e 13 ms, indicando alta responsividade.
- **Throughput:** Superior a 110 req/s em cenários de ataque, demonstrando escalabilidade e capacidade de processamento.

Esses resultados garantem uma experiência de usuário fluida e posicionam o sistema para ambientes de produção exigentes, com grande capacidade de escalar para aplicações críticas, como:

- **APIs de segurança intermediária:** Proteção em tempo real entre usuários e LLMs.
- **Camadas de validação de entrada:** Segurança das requisições antes do modelo.
- **Gateways de conteúdo gerado por LLM:** Filtragem e garantia de qualidade do output.

O desempenho otimizado em cenários maliciosos é atribuído ao bloqueio precoce de prompts inválidos, que demandam **menos processamento profundo** e resultam em **respostas mais rápidas**.

Conclusão da Discussão

Os resultados deste estudo demonstram a eficácia do **Chat Seguro para LLMs**. As principais conclusões são:

- **Excelente Capacidade de Detecção:** O sistema identifica ameaças de forma eficaz, especialmente contra ataques de Prompt Injection.
- **Baixa Incidência de Falsos Positivos:** Garante alta confiabilidade e evita interrupções desnecessárias.
- **Desempenho Satisfatório:** Com baixa latência e alto throughput, é adequado para cenários de produção em larga escala.

Recomenda-se apenas pequenos ajustes para otimizar a precisão em prompts legítimos. Em suma, o Chat Seguro para LLMs é **efetivo, rápido e robusto**, cumprindo o objetivo de mitigar riscos de segurança e proteger aplicações de LLMs.