Prompt injection-aanvallen

VIBE chatbot

Groep 3

Milan Frees

Xander Thijs

Senne Reekmans

Abdulrahman Akil

Projectomschrijving

Dit onderzoek focust op het vinden van de meest effectieve Natural Language Processing (NLP)-technieken om prompt injection-aanvallen op te sporen en tegen te gaan in VIBE (Virtual Intelligent Business Executive), een AI-gestuurde virtuele assistent. Het hoofddoel is om een beveiligingssysteem te ontwikkelen dat de betrouwbaarheid, veiligheid en privacy van het systeem garandeert, terwijl het ook een prettige gebruikerservaring biedt voor bezoekers van de Corda Arena. Het gebruik van AI-systemen zoals VIBE brengt natuurlijk risico's met zich mee op het gebied van cybersecurity en privacy, vooral in een omgeving waar gevoelige informatie wordt uitgewisseld. Prompt injection-aanvallen zijn hierbij een groot probleem: gebruikers met slechte intenties kunnen de AI zo manipuleren dat deze niet toegestane acties uitvoert of gevoelige data lekt. Dit kan leiden tot ernstige privacy-schendingen en beveiligingsproblemen. Het is daarom cruciaal dat de privacy van gebruikers altijd beschermd blijft. Bovendien, als de chatbot kwaadaardige vragen niet herkent, kunnen gebruikers misbruik maken van prompts om bijvoorbeeld achterliggende systemen, zoals een database, aan te vallen.

Om deze problemen aan te pakken, wordt in dit onderzoek een vergelijking gemaakt tussen verschillende NLP-technieken. Deze technieken worden grondig onderzocht en tegen elkaar afgewogen, waarbij zowel hun voordelen als hun beperkingen in kaart worden gebracht. Door deze aanpak hopen we inzicht te krijgen in de meest geschikte methoden om prompt injection-aanvallen te detecteren en te voorkomen. Uiteindelijk moet dit leiden tot een sterker beveiligingssysteem dat de privacy en veiligheid van AI-systemen zoals VIBE beter beschermt.

Problem statement

VIBE, een AI-gestuurde virtuele assistent, moet veilig worden geïntegreerd als eerste aanspreekpunt voor bezoekers van de Corda Arena, met aandacht voor privacy en cybersecurity. Hierbij ontstaan enkele uitdagingen door de risico's van AI-systemen in een omgeving waar gevoelige gebruikersdata wordt verwerkt. Gebruikers met slechte intenties kunnen via prompt injection-aanvallen de AI manipuleren om ongeautoriseerde acties uit te voeren, zoals het lekken van persoonlijke gegevens (bijvoorbeeld betalingsinformatie of account details) of het omzeilen van toegangscontroles. Dit vormt een directe bedreiging voor de privacy van gebruikers en de integriteit van achterliggende systemen, zoals databases of API's. Als de chatbot er niet in slaagt om kwaadaardige prompts te detecteren, dan leidt dit tot grote cybersecurity problemen, waaronder denial-of-service-aanvallen (DoS) of het injecteren van schadelijke code. Om het vertrouwen van gebruikers te behouden moet VIBE sterke beveiligingsmaatregelen integreren tegen dergelijke exploits.

Onderzoeksvraag

- Welke techniek van Natural Language Processing (NLP) is het beste in het detecteren en blokkeren van prompt injection-aanvallen in VIBE?

Subvragen

- Wat zijn prompt injection-aanvallen en welke vormen bestaan er?

- Hoe werkt NLP-gebaseerde input sanitatie en welke technieken worden hierbij gebruikt?

- Welke datasets en evaluatiemethoden kunnen worden gebruikt om de effectiviteit van NLP-gebaseerde input sanitatie te testen?

- Hoe accuraat is NLP-gebaseerde input sanitatie bij het detecteren en blokkeren van verschillende typen prompt injection-aanvallen?

- Wat zijn de beperkingen en mogelijke zwakke plekken van NLP-gebaseerde input sanitatie in VIBE?

- Hoe beïnvloedt het gebruik van NLP-gebaseerde input sanitatie de systeemprestaties van VIBE?

Inhoudsopgave

Lijst van gebruikte figuren

Lijst van gebruikte tabellen

Lijst van gebruikte afkortingen

NLP	Natural Language Processing
VIBE	Virtual Intelligent Business Executive

1 Onderzoeksvraag en hypothese

- **Onderzoeksvraag:** Welke techniek van Natural Language Processing (NLP) is het beste in het detecteren en blokkeren van prompt injection-aanvallen in VIBE?

- **Hypothese:** Er bestaat een groot assortiment aan NLP technieken, één van deze gaat zeker voldoen aan de vereisten van VIBE. Een NLP-gebaseerde input sanitatie techniek die gebruik maakt van contextuele analyse en patroonherkenning zal effectiever zijn in het detecteren en blokkeren van prompt injection-aanvallen in VIBE dan eenvoudige keywordfiltering.

2 Onderzoeksmethode

Om de onderzoeksvraag te beantwoorden, start het onderzoek met een literatuurstudie naar prompt injection-aanvallen en bestaande NLP-gebaseerde input sanitatie technieken. Hiermee wordt inzicht verkregen in verschillende methoden en hun effectiviteit. Vervolgens worden de meest relevante technieken geselecteerd op basis van criteria zoals patroonherkenning, contextuele analyse en prestaties.

Na deze selectie worden testdatasets samengesteld met diverse vormen van prompt injection-aanvallen. Deze datasets worden gebruikt om de effectiviteit van de gekozen technieken te evalueren. De tests worden uitgevoerd in een gesimuleerde omgeving, waarbij de nauwkeurigheid en prestaties van de technieken worden gemeten en vergeleken.

Daarnaast wordt een analyse uitgevoerd naar de impact van NLP-gebaseerde input sanitatie op de systeemprestaties van VIBE. Dit omvat metingen van verwerkingssnelheid en resourcegebruik. Tot slot worden de resultaten geanalyseerd en vergeleken om de meest geschikte techniek te identificeren en aanbevelingen te formuleren voor de implementatie hiervan binnen VIBE.

3 Literatuurstudie

4 Uitvoering

4.1 Hoofdstuk

4.1.1 Hoofdstuk

4.2 Hoofdstuk

5 Conclusie

Bibliografie

Bijlagen

A. **Omschrijving Bijlage A**

B. **Omschrijving Bijlage B**

C. **Omschrijving Bijlage C**

A. Omschrijving Bijlage A

B. Omschrijving Bijlage B

C. Omschrijving Bijlage C