ПРЕДЛОЖЕНИЕ

за провеждане на докторантура в редовна форма на обучение
по реда на чл. 21, ал. 7 от ЗВО

в Института по математика и информатика при Българска академия на науките

професионално направление **4.6. Информатика и компютърни науки**
докторска програма „**Информатика**“

научноизследователска област:
Б. Приложения с изкуствен интелект, ориентирани към анализ и обработка на данни
(Подобряване на интерпретируемостта и устойчивостта на ИИ модели)

тема на докторантурата: **Бейсов самообучаващ се модел за вероятностна устойчивост, статистическа калибрация и надеждна оценка на киберзаплахи и уязвимости в контекстуално обусловени големи езикови модели**

научни ръководители: **доц. д.н. Цветелин Заевски и доц. д-р Красимира Иванова**

Актуалност

Центровете за операции по сигурност (SOC) се сблъскват с огромен обем от софтуерни уязвимости, докладвани ежедневно чрез източници като Националната база данни за уязвимости (NVD), както и тяхната прогнозируема експлоатация спрямо Exploit Prediction Scoring System (EPSS). Тези уязвимости, известни като Common Vulnerabilities and Exposures (CVE) се различават значително по тежест, експлоатируемост и въздействие върху бизнеса, което прави интелигентното приоритизиране от съществено значение. В контекста на нарастваща автоматизация, използването на големи езикови модели (LLM) за анализ, прогнозиране на експлоатацията и приоритизацията на такива заплахи става все по-съществено. Въпреки това, липсата на механизми за обоснована оценка на несигурност, прогнозируемост, обяснимост и надеждност ограничава приложимостта на тези модели в чувствителни среди.

Цел на докторантурата

Това изследване има за цел да разработи Бейсов модел за машинно самообучение, който подобрява вероятностната устойчивост, статистическата калибрация и надеждното вземане на решения в контекстуално обусловени големи езикови модели, прилагани за приоритизация на заплахи от уязвимости (CVE). Целта не е изграждането на пълна продукционна система, а теоретично моделиране и разработване на потвърждение на концепцията (PoC), демонстриращо как интелигентни агенти, осъзнаващи своята несигурност, могат да подпомогнат прогнозируемостта на експлоатация, както и управлението на уязвимости.

Научни задачи и методи за изпълнение

В рамките на дисертационното изследване се предвижда изпълнение на следните научни задачи:

1. **Разработване на Бейсов модел за оценка на несигурността в големи езикови модели (LLM) –** Ще се изследва как Бейсовото заключение може да бъде

приложено към LLM при приоритизиране на уязвимости от тип CVE. Ще се използват техники като Monte Carlo dropout за апроксимация на несигурността.

2. **Изследване и прилагане на методи за статистическа калибрация на изходните прогнози на LLM –** Ще бъдат тествани различни подходи за калибриране, с цел подобряване на надеждността и доверието в прогнозите, особено в условия на несигурност.
3. **Интегриране на контекстуални фактори в процеса на вземане на решения –** Ще се моделират и включат сигнали като критичност на актив, CVSS, EPSS стойности, наличност на експлойти и други контекстуални метаданни за по-прецизна оценка на риска.
4. **Разработка и експериментиране с прагове за вземане на решения –** Ще се изследват и дефинират прагови стойности на достоверност, които да управляват автоматизираното или делегирано вземане на решения от страна на LLM агентите, като ще се използва референтната рамка на EPSS за сравнение.
5. **Дизайн и валидиране на PoC система за приоритизация на уязвимости с оценка на несигурността –** Ще бъде изградена PoC система, използваща модели като DistilBERT и RoBERTa, обработваща публични набори от данни (NVD, ExploitDB, MITRE ATT&CK), и оценяваща устойчивостта, прогнозируемостта и интерпретируемостта и ефективността на предложения модел.

Очаквани резултати

- Теоретична математическа Бейсова рамка за оценка на несигурността в LLM.
- Калибрационни методи за повишаване на доверието в прогнозите на LLM.
- Теоретичен модел за интеграция на контекстуални сигнали при оценка на риска.
- Дефинирани прагови стойности и стратегии за делегиране на решения при несигурност.
- PoC система, валидирана с публични CVE данни и LLM модели.

Въздействие

Изследването ще допринесе за повишаване на сигурността и надеждността в автоматизирани решения в киберпространството чрез въвеждане на методи за управление на несигурността в големи езикови модели. То ще предостави приложими подходи за изграждане на доверими системи, съвместими с етични и регулаторни изисквания. Резултатите имат потенциал за широко въздействие в сфери като критична инфраструктура, финанси, здравеопазване и други, където доверието в решения, вземани с участието на интелигентни алгоритми, е от ключово значение.

Място на зачисляване

Секция „Софтуерни технологии и информационни системи“ (ИМИ-БАН).
Поради интердисциплинарността на работата плановете за развитие и резултатите ще бъдат дискутирани и на семинари на секция „Изследване на операциите, вероятности и статистика“.

Използвана научна инфраструктура

Хемус