VEĽKÉ JAZYKOVÉ MODELY (LLMs) PRE BEZPEČNOSŤ KONCOVÝCH ZARIADENÍ

UNIVERZITA PAVLA JOZEFA ŠAFÁRIKA V KOŠICIACH

Financované Európskou úniou NextGenerationEU

PLÁN OBNOVY

MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Projekt Kompetenčné centrum kybernetickej bezpečnosti na UPJŠ (KCKB UPJŠ) financovaný Európskou úniou z prostriedkov Plánu obnovy a odolnosti Slovenskej republiky, kód projektu: 17R05-04-V01-00007.

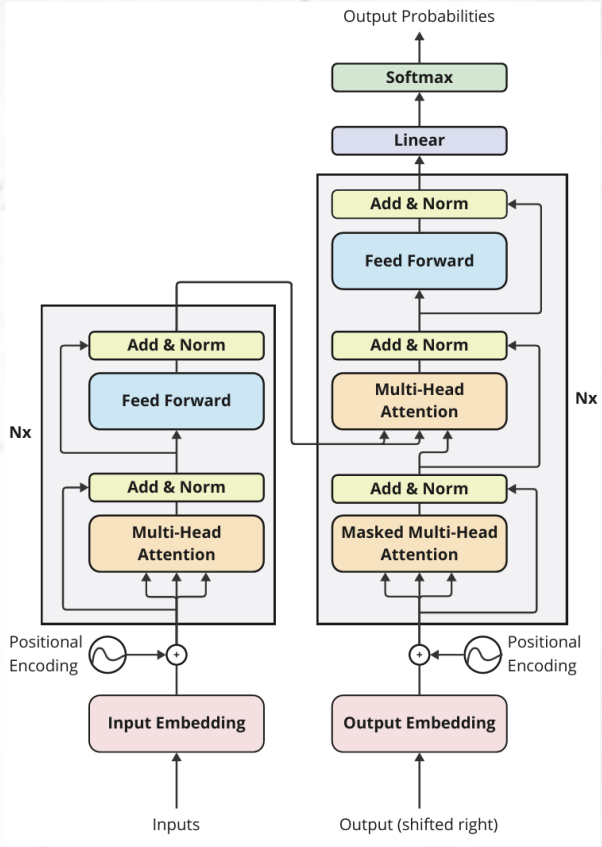
Analýza malvéru

- Dôležitou súčasťou bezpečnosti koncových zariadení je **analýza malvéru** – umožňuje odhaľovať škodlivé súbory, podozrivé procesy a nežiaduce správanie programov
- Malvér je čoraz rozmanitejší, sofistikovanejší a ťažšie detekovateľný
- Analýza malvéru sa delí na tri prístupy: statickú (bez spustenia programu), dynamickú (vyžaduje spustenie programu v izolovanom virtuálnom prostredí) a hybridnú (kombinácia prístupov)
- Tradičné prístupy analýzy malvéru majú obmedzenia pri nových a obfuskovaných hrozbách
- Čoraz viac sa uplatňujú metódy založené na umelej inteligencii a hlbokom učení
- Medzi perspektívne architektúry hlbokého učenia patria **Transformery** – dokážu zachytávať širší kontext a zložité vzory v dátach

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Transformery

- Transformer využíva mechanizmus pozornosti, vďaka ktorému dokáže pracovať s kontextom celej vstupnej sekvencie (napr. tokenov)
- Štandardná architektúra Transformeru pozostáva z dvoch podsietí: encoder a decoder
- **Encoder** mapuje vstupnú sekvenciu na kontextové reprezentácie, ktoré **decoder** ďalej spracúva pri generovaní výstupu
- V porovnaní s rekurentnými modelmi umožňuje efektívnejšie spracovanie sekvenčných dát a lepšiu paralelizáciu výpočtov



Zdroj: A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Predtrénované Transformery

- Predtrénované Transformery sa najprv učia všeobecné reprezentácie z veľkého objemu dát a následne sa dolaďujú na konkrétnu úlohu
- Predtrénovanie sa často realizuje pomocou self-supervised learning – učenie z neoznačených dát

Základné typy predtrénovaných Transformerov:

- BERT – porozumenie textu, klasifikácia
- CANINE – znaková reprezentácia bez tokenizácie
- ViT – spracovanie obrazu
- GPT-2 – generovanie textu

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Transformery v analýze malvéru

- Uplatňujú sa pri práci s rôznymi typmi vstupov, napr. s binárnymi sekvenciami, assemblerom, API volaniami, grafmi, obrazom a sieťovým správaním
- Vedia zachytávať rôzne typy vzťahov v dátach, napr. sekvenčné, sémantické, štrukturálne, priestorové a temporálne korelácie

Možnosti využitia Transformerov v analýze malvéru:

- Detekcia malvéru – rozlišovanie medzi škodlivými a benígnymi vzorkami
- Klasifikácia malvéru – priraďovanie vzoriek k rodinám alebo triedam malvéru
- Detekcia podobnosti binárneho kódu – porovnávanie binárnych vzoriek a hľadanie príbuzných alebo odvodených variantov
- Vysvetľovanie rozhodnutí modelov – interpretácia, prečo model označil vzorku ako škodlivú
- Analýza techník obchádzania detekcie – skúmanie, ako sa malvér snaží vyhnúť jeho odhaleniu

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Prehľad výskumu

Financované Európskou úniou
NextGenerationEU

PLÁN OBNOVY

MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Efficient Malware Analysis Using Metric Embeddings

E. M. Rudd, D. Krisiloff, S. Coull, D. Olszewski, E. Raff, J. Holt (2024)

Ciele:

- Navrhnúť kompaktnú reprezentáciu malvéru, ktorá zjednoduší jeho analýzu a bude využiteľná vo viacerých úlohách (napr. detekcia malvéru, klasifikácia rodín a určovanie atribútov malvéru)
- Zachytiť nielen statické vlastnosti PE súborov, ale aj informácie o ich funkcionalitách
- Znížiť výpočtové a pamäťové nároky oproti práci s pôvodnou vysokorozmernou reprezentáciou

Metodológia:

- PE súbory sú najprv opísané pomocou statických príznakov
- Následne sa prevádzajú do nízkorozmerných embeddingov pomocou neurónovej siete
- Pri učení sa využívajú aj informácie z nástroja CAPA, ktorý zisťuje aké funkcionality obsahuje analyzovaný súbor
- Autori porovnávajú tri prístupy učenia: contrastive loss, Spearman loss a ich kombináciu

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Efficient Malware Analysis Using Metric Embeddings

E. M. Rudd, D. Krisiloff, S. Coull, D. Olszewski, E. Raff, J. Holt (2024)

Výsledky:

- Embeddingy sa osvedčili pri detekcii malvéru, klasifikácii rodín aj určovaní atribútov malvéru
- Najlepšie výsledky priniesla kombinácia oboch stratových funkcií
- Výkon embeddingov zostal blízky referenčným modelom, hoci reprezentácia bola výrazne menšia
- 32-rozmerná reprezentácia výrazne znížila nároky na úložisko aj výpočtové zdroje

Prínos:

- Článok ukazuje, že malvér možno reprezentovať jednoduchšie a úspornejšie, bez veľkej straty kvality
- Navrhnuté embeddingy môžu slúžiť ako univerzálna reprezentácia pre viacero analytických úloh

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Malware Detection for Portable Executables Using a Multi-input Transformer-based Approach

T.-L. Huoh, T. Miskell, O. Barut, Y. Luo, P. Li, T. Zhang (2024)

Ciele:

- Navrhnúť nový prístup na detekciu škodlivých PE súborov využívajúci raw bajty PE hlavičiek namiesto navrhnutých príznakov
- Overiť, či kombinácia CNN a Transformer architektúr dokáže lepšie zachytiť priestorové aj sekvenčné vlastnosti dát
- Preskúmať prínos multi-input prístupu, v ktorom sa jednotlivé časti PE hlavičky učia nezávisle

Metodológia:

- Ako vstup použili dve sekvencie raw bajtov získané z PE súboru (file header, optional header)
- Navrhli tri experimentálne varianty: (Study 1) CNN a early fusion, (Study 2) CNN + Transformer a early fusion, (Study 3) CNN + Transformer a late fusion
- Model testovali na dvoch verejných datasetoch DeepDetectNet (DDN) a Ransomary, ako referenčné prístupy použili LightGBM a MalConv

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Malware Detection for Portable Executables Using a Multi-input Transformer-based Approach

T.-L. Huoh, T. Miskell, O. Barut, Y. Luo, P. Li, T. Zhang (2024)

Výsledky:

- Vo všetkých porovnaniach dosiahol najlepšie výsledky model Study 3, teda model s kombináciou CNN, Transformerov a late fusion stratégie
- Inferenčný čas navrhovaného modelu bol mierne vyšší než pri LightGBM, ale výrazne nižší než pri MalConv

Prínos:

- Článok ukazuje, že raw bajty PE hlavičiek môžu byť dostatočne informatívnym vstupom pre presnú detekciu malvéru
- Prínosom je aj návrh multi-input Transformer-based architektúry, ktorá umožňuje samostatné učenie z rôznych častí PE štruktúry
- Výsledky naznačujú, že late fusion a kombinácia CNN s Transformermi vedú k lepšej reprezentácii dát než jednoduchšie referenčné prístupy

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Nebula: Self-Attention for Dynamic Malware Analysis

D. Trizna, L. Demetrio, B. Biggio, F. Roli (2024)

Ciele:

- Navrhnúť univerzálny model na dynamickú analýzu malvéru, ktorý dokáže spracovať rôzne typy behaviorálnych údajov z dynamických logov
- Prekonať obmedzenia predchádzajúcich prístupov, ktoré sa zameriavali len na homogénne vstupy, napr. API volania, a neuvažovali ďalšie dôležité informácie, ako sieťové a súborové operácie
- Overiť, či architektúra založená na self-attention dokáže zlepšiť detekciu a klasifikáciu malvéru, najmä pri nízkej miere falošne pozitívnych výsledkov

Metodológia:

- Autori navrhli architektúru Nebula – Transformer-based model určený na spracovanie dynamických logov vzniknutých pri spustení programu v kontrolovanom prostredí
- Súčasťou prístupu je viacero krokov predspracovania (tokenizácia, filtrovanie, normalizácia a kódovanie dát)
- Model bol testovaný na úlohách detekcie malvéru aj klasifikácie do rodín na troch datasetoch
- Autori skúmali aj self-supervised pre-training a interpretáciu modelu pomocou explainable AI techník

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Nebula: Self-Attention for Dynamic Malware Analysis

D. Trizna, L. Demetrio, B. Biggio, F. Roli (2024)

Výsledky:

- Model Nebula v priemere prekonal predchádzajúce prístupy, pričom pri nízkej miere falošne pozitívnych výsledkov dosiahol zlepšenie až o 12%
- Ukázalo sa, že navrhnutý prístup dokáže účinne kombinovať rôzne typy behaviorálnych informácií
- Self-supervised pre-training dosiahol porovnateľný výkon s plne supervidovanými modelmi už pri použití iba 20% tréningových dát
- Ukázalo sa, že model venuje zvýšenú pozornosť tým častiam vstupu, ktoré súvisia so škodlivými aktivitami a charakteristickými znakmi malvérových rodín

Prínos:

- Článok prináša univerzálny Transformer-based prístup na dynamickú analýzu malvéru, ktorý dokáže pracovať s rôznorodými dátami
- Praktický prínos spočíva v tom, že model dosahuje dobrý výkon aj pri menšom množstve označených dát, a zároveň poskytuje určitú mieru interpretovateľnosti

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Transformer-Based API Call Sequence Modeling for Dynamic Malware Detection

N. Youssef, N. Elbaraway, A. Elmaghraby (2025)

Ciele:

- Navrhnúť prístup na dynamickú detekciu a klasifikáciu malvéru založený na modelovaní sekvencií API volaní
- Overiť, či BERT umožní lepšie zachytiť vzťahy v usporiadaných sekvenciách API volaní než bežne používané prístupy založené na CNN a LSTM
- Overiť, či kombinácia extrakcie príznakov neurónovou sieťou a transformerového modelovania sekvencií zlepší detekciu aj pri neúplných execution logoch

Metodológia:

- Autori vychádzajú z dynamických execution logov, ktoré zachytávajú správanie programu počas jeho spustenia v emulovanom prostredí
- Použijú neurónovú sieť na extrakciu príznakov, ktorá zachytáva dôležité charakteristiky správania malvéru
- Takto získané reprezentácie potom spracuje BERT-based klasifikátor, ktorý modeluje usporiadané sekvencie API volaní

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Transformer-Based API Call Sequence Modeling for Dynamic Malware Detection

N. Youssef, N. Elbaraway, A. Elmaghraby (2025)

Výsledky:

- V porovnaní s prístupmi Neurlux, Nebula a Quo Vadis dosiahol model vyššie F1-skóre vo viacerých triedach, najmä pri clean súboroch a ransomware
- Výsledky zároveň naznačujú, že zachovanie poradia API volaní podporuje robustnosť modelu aj pri neúplných execution logoch

Prínos:

- Článok ukazuje, že transformerové modelovanie sekvencií API volaní je vhodné na dynamickú analýzu malvéru
- Prínosom je hybridný prístup, ktorý spája extrakciu príznakov neurónovou sieťou s kontextovým modelovaním pomocou BERT
- Práca naznačuje potenciál rozšírenia o ďalšie behaviorálne indikátory, napr. sieťovú aktivitu, pamäťové operácie a interakcie so súborovým systémom

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Transformers for End-to-End InfoSec Tasks: A Feasibility Study

E. M. Rudd, M. S. Rahman, P. Tully (2022)

Ciele:

- Preskúmať, či sú Transformery použiteľné na end-to-end úlohy v informačnej bezpečnosti (pri učení priamo z raw dát bez extrakcie príznakov)
- Overiť ich využiteľnosť na dvoch odlišných typoch dát – URL adresách a PE súboroch
- Zistiť, aké úpravy modelu a tréningové stratégie sú potrebné, aby Transformery dosiahli konkurencieschopný výkon pri detekčných úlohách

Metodológia:

- Autori navrhli end-to-end transformerové modely, ktoré pracujú priamo s raw vstupmi
- Pre URL adresy porovnali viacero tréningových režimov: klasifikáciu bez predtrénovania, auto-regresívne predtrénovanie a mixed objective training, ktorý kombinuje klasifikačný cieľ s cieľom predikcie ďalšieho znaku
- Pre PE súbory navrhli Byte Transformer s adaptive attention span, aby bolo možné spracovať dlhé sekvencie bajtov efektívnejšie z hľadiska pamäte
- Výsledky porovnávali s viacerými referenčnými modelmi

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Transformers for End-to-End InfoSec Tasks: A Feasibility Study

E. M. Rudd, M. S. Rahman, P. Tully (2022)

Výsledky:

- Pri URL adresách predtrénovanie neprinieslo zlepšenie, avšak mixed objective training výkon zlepšil
- Navrhnutý URL Transformer dosiahol porovnateľný výkon s najlepšími referenčnými modelmi
- Pri PE súboroch dosiahol Byte Transformer porovnateľný výkon s modelom MalConv, ale zaostával za modelmi LightGBM a MG-CNN
- Jednoduché techniky kompresie a kódovania bajtov neviedli k ďalšiemu zlepšeniu výkonu

Prínos:

- Článok ukazuje, že Transformery možno použiť aj pri end-to-end úlohách v informačnej bezpečnosti
- Autori navrhli dynamicky vyvažovanú kombinovanú stratovú funkciu, ktorá pri URL klasifikácii zlepšila výsledky
- Práca zároveň definuje limity Transformerov pri dlhých bajtových sekvenciách, najmä z hľadiska škálovateľnosti a výpočtovej náročnosti

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Unveiling the veiled: An early stage detection of fileless malware

N. Singh, S. Tripathy (2025)

Ciele:

- Navrhnúť prístup na včasnú detekciu bezsúborového malvéru, keďže existujúce prístupy ho často zachytávajú až v neskorších štádiách útoku
- Detegovať bezsúborové útoky pred ich plne operačnou fázou, aby sa znížilo riziko poškodenia systému a úniku dát
- Využiť analýzu pamäte a väzbu na rámec MITRE ATT&CK na identifikáciu štádia bezsúborového útoku

Metodológia:

- Autori navrhli rámec Argus, ktorý v reálnom čase sleduje systém a vyhľadáva podozrivé procesy
- Z ich pamäťových výpisov extrahuje kľúčové príznaky správania
- Pomocou modelu Llama generuje textové vysvetlenia extrahovaných príznakov
- Tieto vysvetlené príznaky sa porovnávajú s datasetom odvodeným z MITRE ATT&CK, pričom BERT model kombinovaný s MLP určuje, v ktorom štádiu sa bezsúborový útok nachádza

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

Unveiling the veiled: An early stage detection of fileless malware

N. Singh, S. Tripathy (2025)

Výsledky:

- Navrhnutý rámec dokázal úspešne detegovať bezsúborový malvér vo veľkej časti analyzovaných vzoriek (96%)
- Značnú časť útokov zachytil už v predoperačnom štádiu
- Ďalšie prípady bezsúborového malvéru boli identifikované v operačnom štádiu
- V porovnaní s doterajšími prístupmi sa ukázalo, že Argus je silný najmä v tom, že bezsúborový malvér zachytáva skôr, nie až v závere útoku

Prínos:

- Autori predstavili rámec Argus na včasnú detekciu bezsúborového malvéru
- Článok ukazuje, že bezsúborový malvér možno vo viacerých prípadoch zachytiť ešte pred operačnou alebo finálnou fázou útoku

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

A Deep Learning-Based Model for Malware Detection Using Transformer Architecture

N. Labied, H. Benbrahim, A. Amine (2024)

Ciele:

- Navrhnúť systém na detekciu malvéru v PE súboroch pre Windows založený na transformerovej architektúre
- Overiť, či Transformery dokážu zachytiť zložité vzory v assembleri a odlíšiť škodlivé a benígne súbory presnejšie než tradičné prístupy

Metodológia:

- Autori analyzujú PE súbory pre Windows pomocou viacstupňového postupu
- Zo súborov najprv získajú assemblerový kód a ďalšie statické príznaky (napr. importované DLL knižnice a textové reťazce)
- Assemblerový kód spracúva Transformer, ktorý sa zameriava na dôležité časti kódu
- Všetky získané informácie sa následne spoja a použijú na binárnu klasifikáciu – rozhodnutie, či ide o malvér alebo benígny súbor

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

A Deep Learning-Based Model for Malware Detection Using Transformer Architecture

N. Labied, H. Benbrahim, A. Amine (2024)

Ciele:

- Navrhnúť systém na detekciu malvéru v PE súboroch systému Windows založený na transformerovej architektúre
- Overiť, či Transformery dokážu zachytiť zložité vzory v assembleri a odlíšiť škodlivé a benígne súbory presnejšie než tradičné prístupy

Metodológia:

- Autori analyzujú PE súbory systému Windows pomocou viacstupňového postupu
- Zo súborov najprv získajú assemblerový kód a ďalšie statické príznaky (napr. importované DLL knižnice a textové reťazce)
- Assemblerový kód spracúva Transformer, ktorý sa zameriava na dôležité časti kódu
- Všetky získané informácie sa následne spoja a použijú na binárnu klasifikáciu – rozhodnutie, či ide o malvér alebo benígny súbor

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

A Deep Learning-Based Model for Malware Detection Using Transformer Architecture

N. Labied, H. Benbrahim, A. Amine (2024)

Výsledky:

- Navrhnutý systém dosiahol vysokú presnosť pri detekcii malvéru v PE súboroch systému Windows, a zároveň vyššiu presnosť než tradičné modely
- Model taktiež umožnil rýchlu analýzu jednotlivých súborov, čo naznačuje jeho praktickú použiteľnosť

Prínos:

- Článok ukazuje, že transformerová architektúra je použiteľná na detekciu malvéru v PE súboroch a dokáže pracovať s komplexnými vzormi v assembleri
- Práca podporuje myšlienku, že Transformery môžu poskytnúť presnejšie a adaptívnejšie riešenia na detekciu moderného malvéru

Financované Európskou úniou NextGenerationEU
PLÁN OBNOVY
MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY

LLMs pre bezpečnosť koncových zariadení

Financované Európskou úniou
NextGenerationEU

PLÁN OBNOVY

MINISTERSTVO INVESTÍCIÍ, REGIONÁLNEHO ROZVOJA A INFORMATIZÁCIE SLOVENSKEJ REPUBLIKY