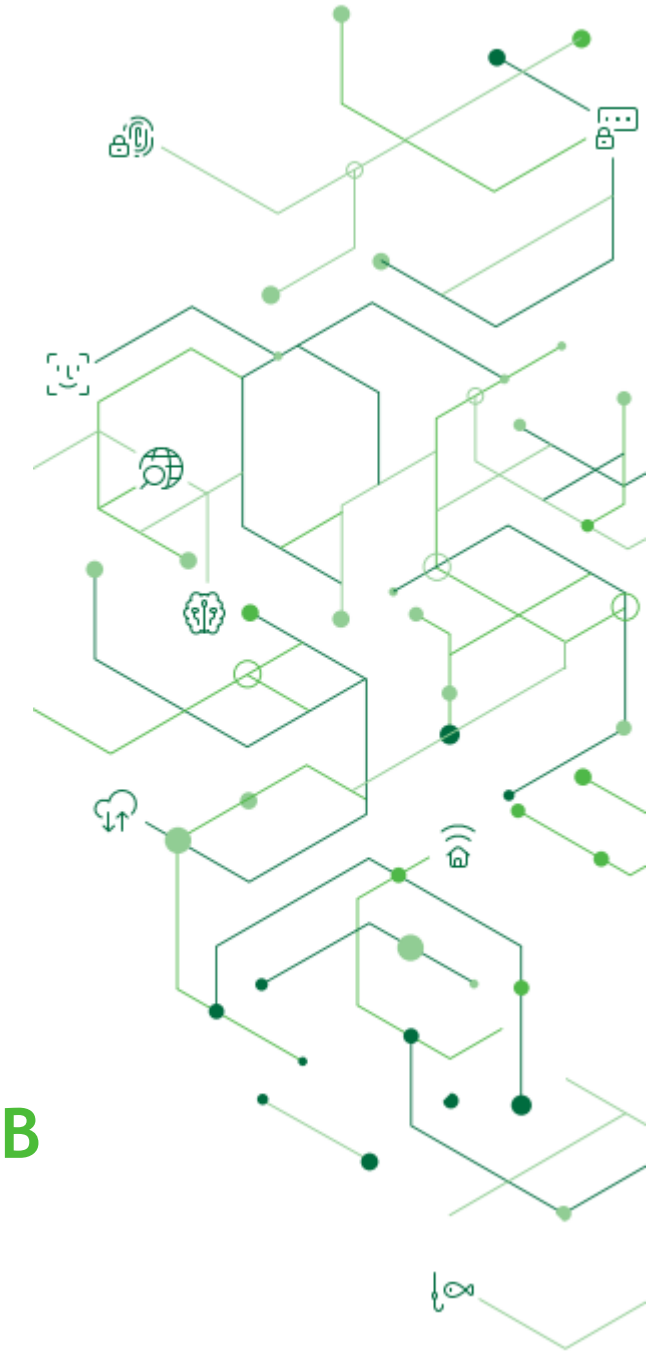
Nagy nyelvi modellek biztonsága



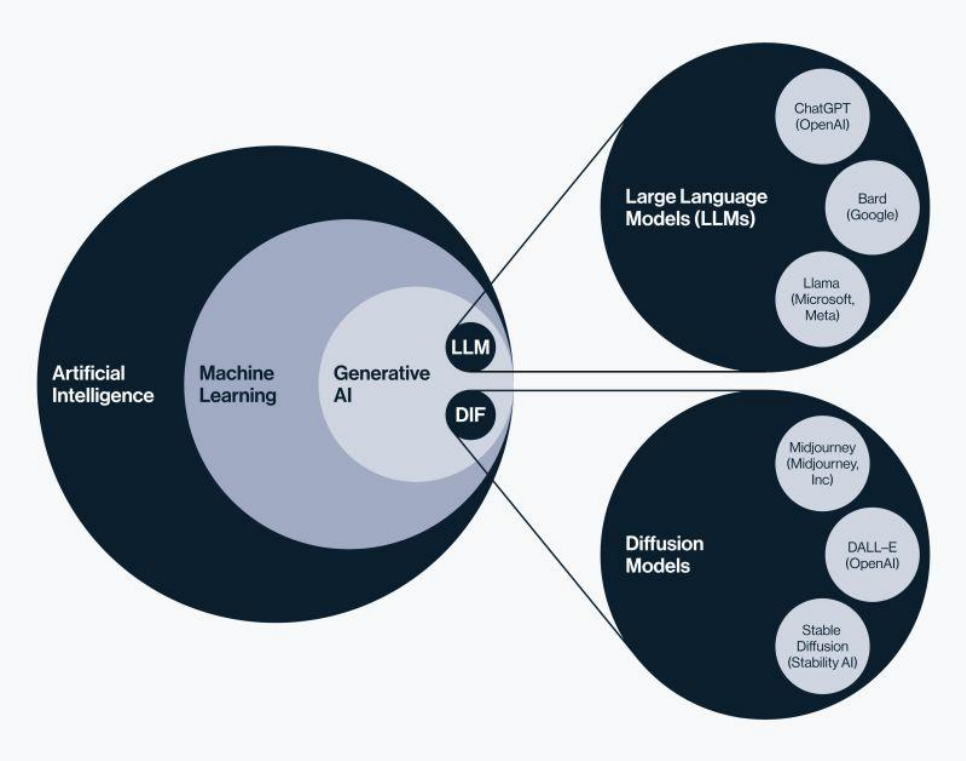
ELTE | IK
INFORMATIKAI KAR

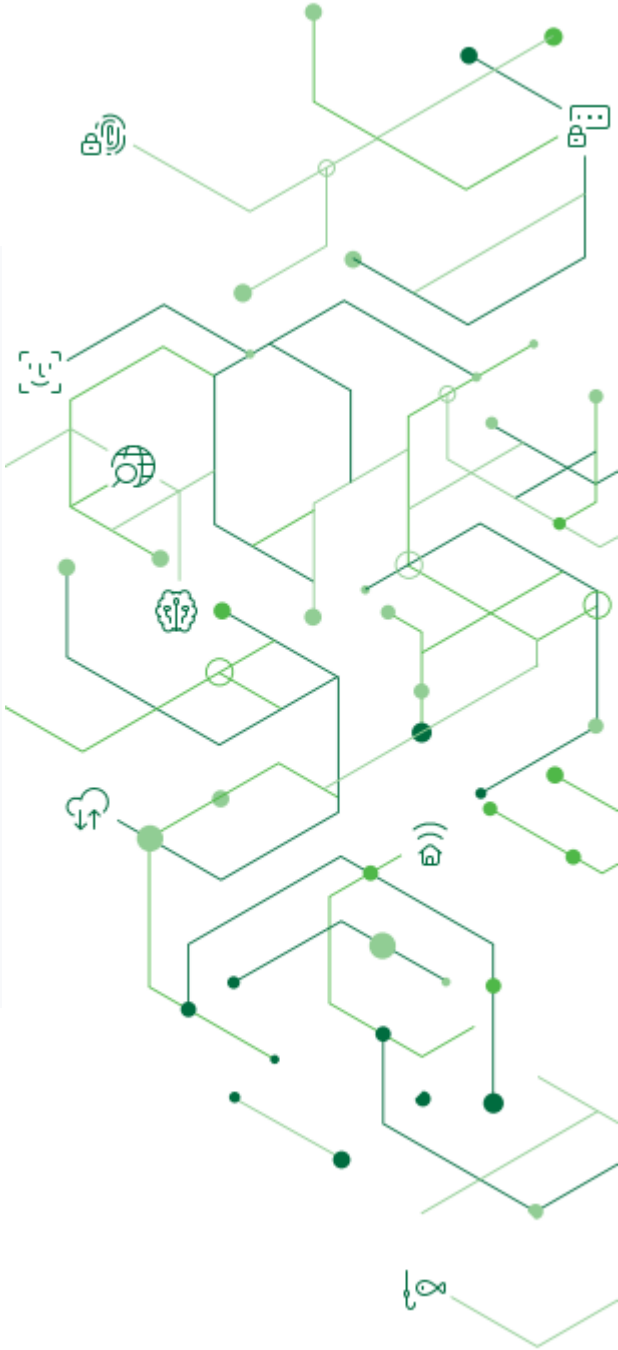
otpbank

KIBERLAB

Mesterséges Intelligencia

- Robbanásszerű fejlődése
- Új kihívások az adatbiztonság területén
- A rendszereket a feladatok komplexitása alapján választjuk ki:
- Egyszerű problémák: gépi tanulási technikák
- Összetettebb feladatok: generatív mesterséges intelligencia, pl. Nagy Nyelvi Modellek (LLM-ek).





Tanulási folyamat

- Első lépésként az OWASP Top 10-es listájának tanulmányozása.
- Kiemelt figyelem a Prompt Injection sebezhetőségre.

LLM01: Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02: Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03: Training Data Poisoning

This occurs when LLM training data is tampered, introducing vulnerabilities or biases that compromise security, effectiveness, or ethical behavior. Sources include Common Crawl, WebText, OpenWebText, & books.

LLM04: Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05: Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre- trained models, and plugins can add vulnerabilities.

LLM06: Sensitive Information Disclosure

LLM's may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. It's crucial to implement data sanitization and strict user policies to mitigate this.

LLM07: Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control. This lack of application control makes them easier to exploit and can result in consequences like remote code execution.

LLM08: Excessive Agency

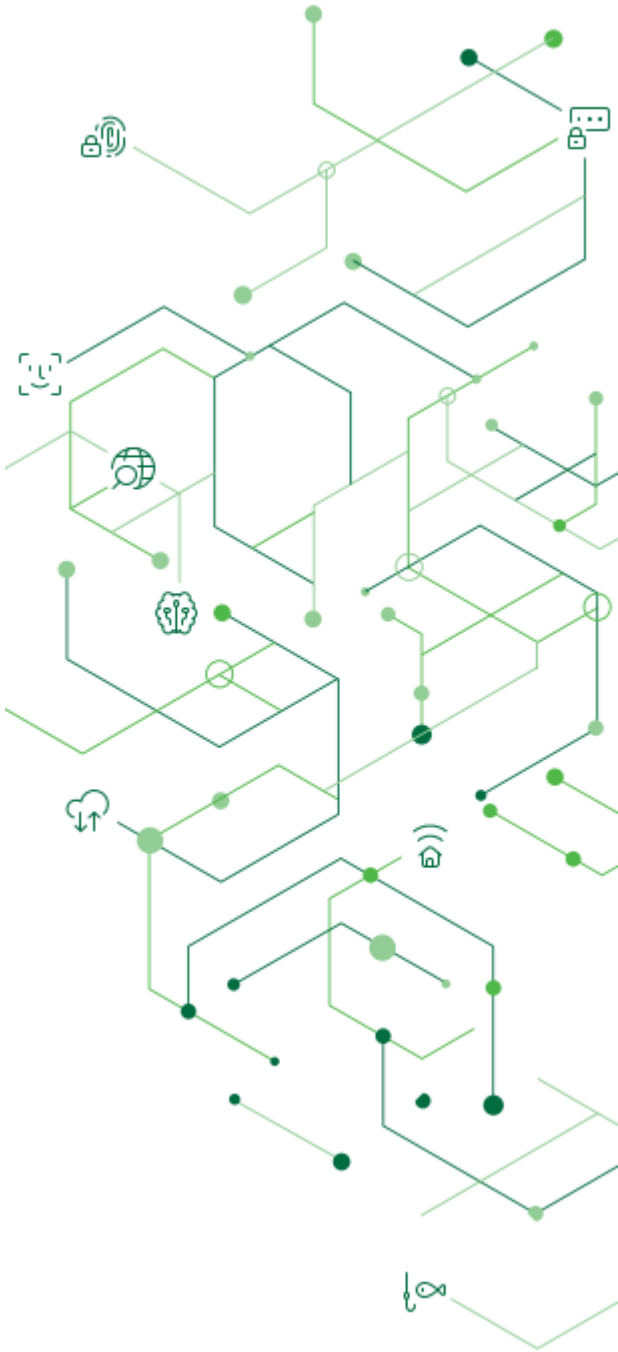
LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09: Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

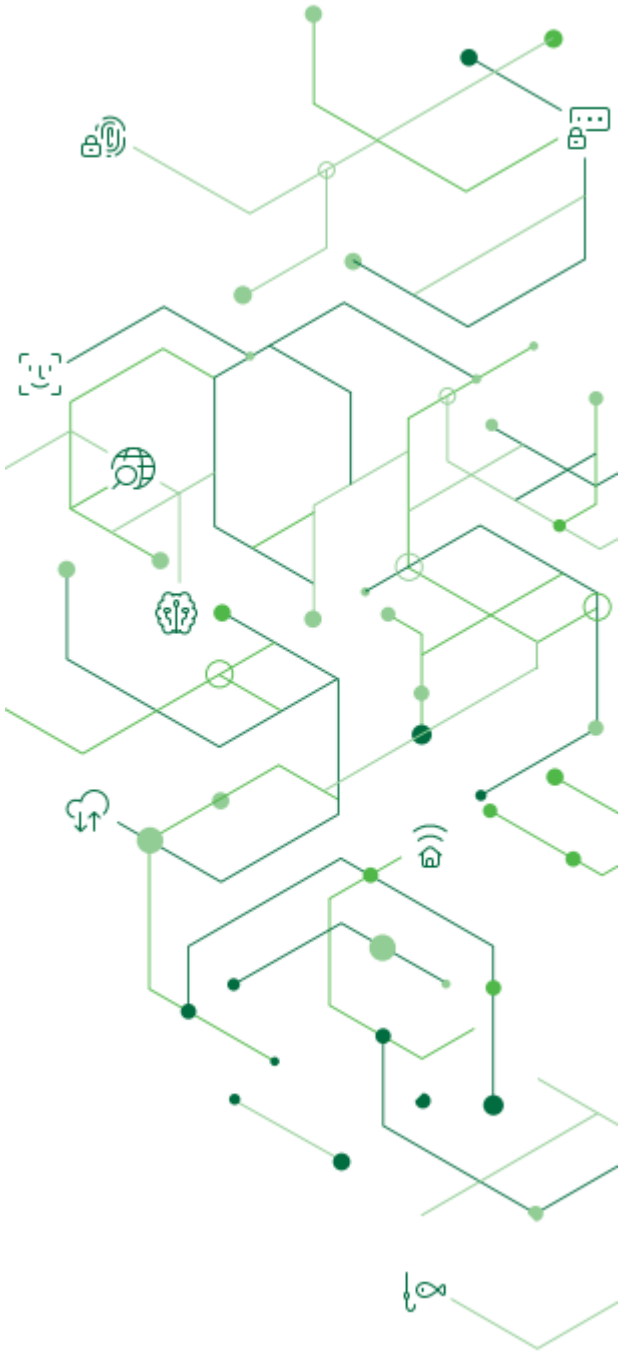
LLM10: Model Theft

This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.



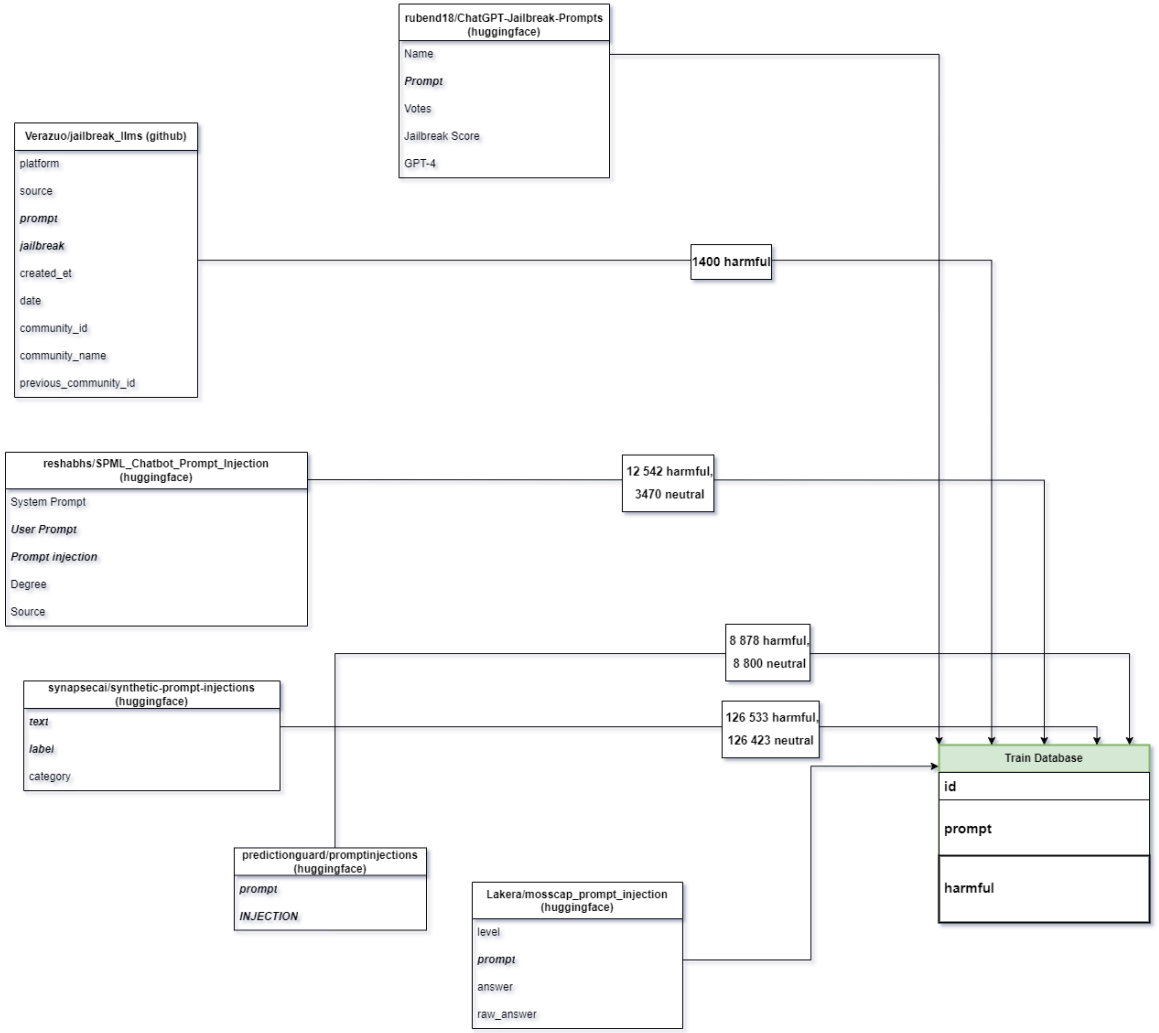
Célok

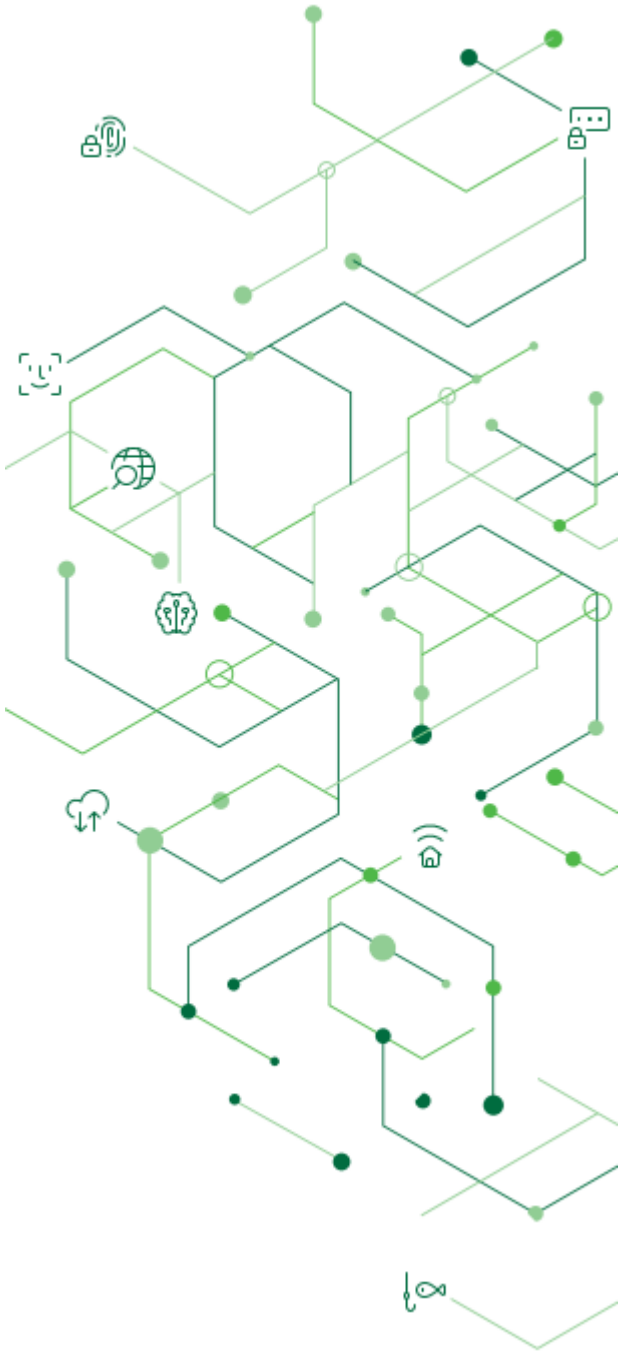
- Prototípus létrehozása, amely hatékonyan szűri és elemzi a felhasználói kéréseket.
- Saját modell kerül a nagy nyelvi modell elé, amely először ellenőrzi az inputokat.
- Kiszűri a potenciálisan káros kéréseket, mielőtt azok a fő modellhez jutnának.
- A szűrőmodell mennyire hatékony a támadások blokkolásában.



Prototípus Kialakítása

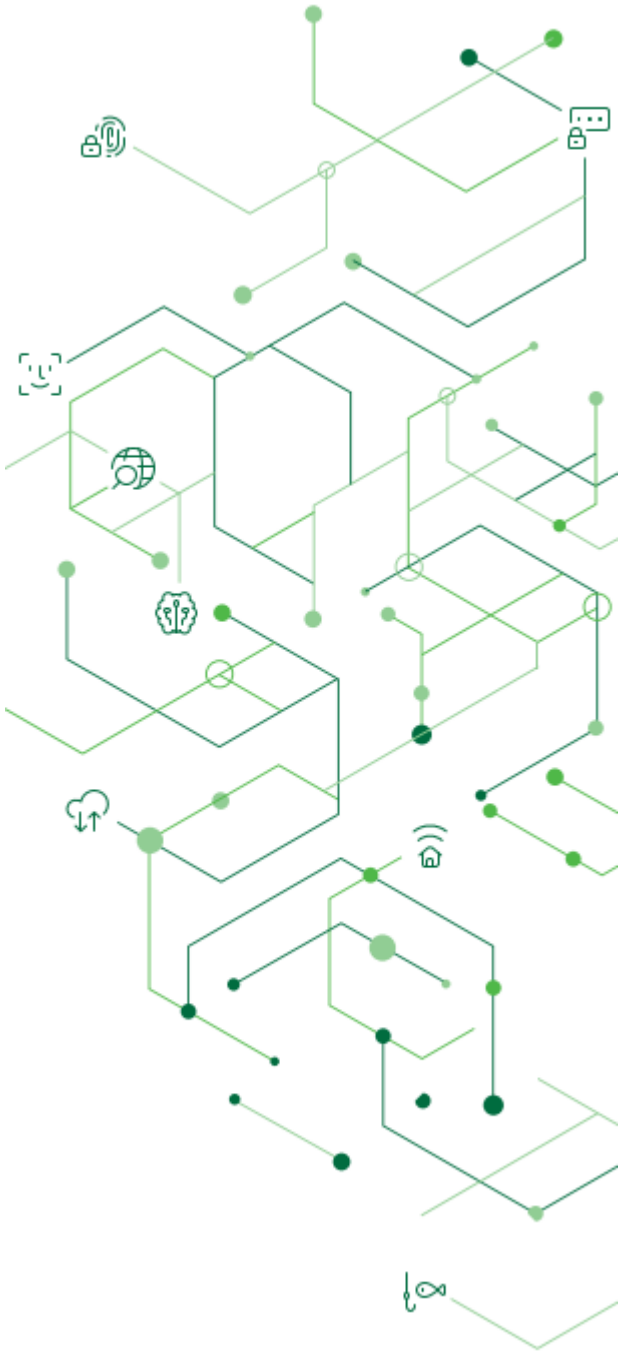
- Prototípus sikeres létrehozása és hatékonyságának felmérése
- Több különböző forrás felhasználása
- Negyedmillió prompt





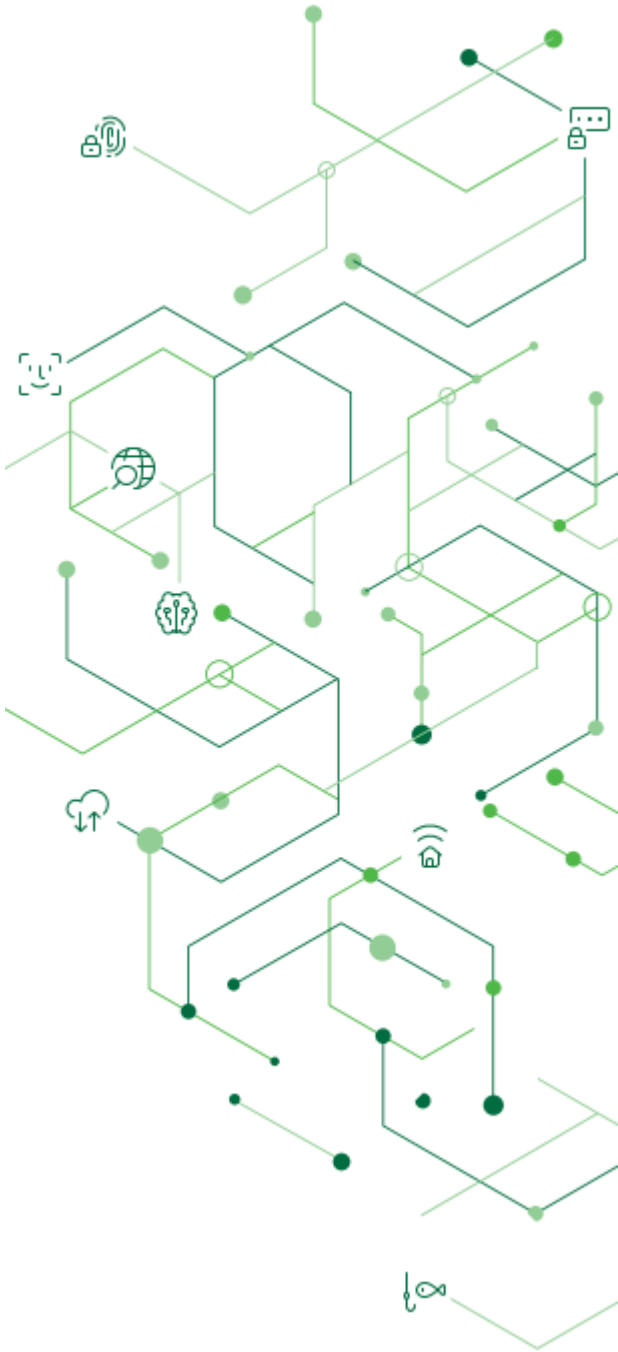
Modell Kiválasztása

- A prototípushoz a **DistilBERT** modellt választottuk, amely a BERT nyelvi modell kisebb és gyorsabb változata.
- A BERT képes a szavakat mindkét irányból értelmezni, biztosítva a pontos kontextuális megértést.
- A DistilBERT optimalizált formája kevesebb számítási erőforrást igényel, de hasonló eredményeket biztosít.
- Ugyanazon a nagy méretű angol nyelvű adathalmazon lett előtanítva, így nyelvi megértési képességei hasonlóak.



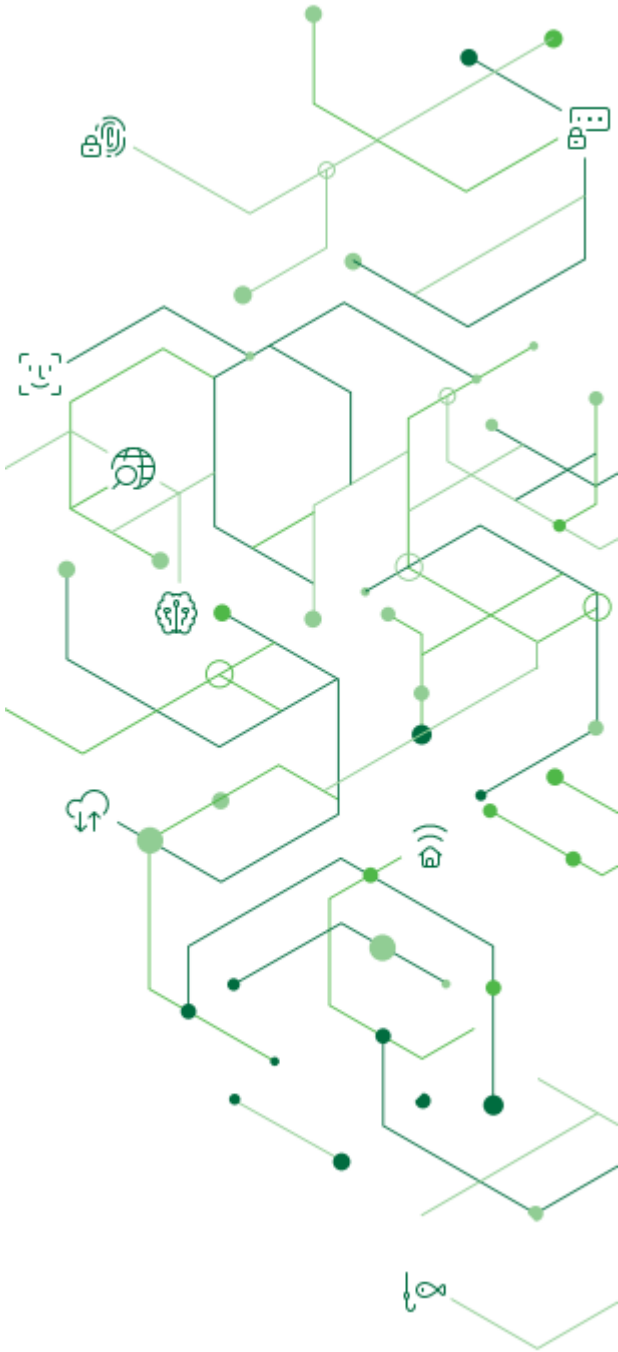
Eredmények

- A prototípus végül **93%-os pontosságot** ért el a támadások felismerésében.



További célok

- Adatbázis frissítése
  - A támadások fejlődése miatt folyamatosan frissíteni kell az adatbázist.
- Integráció a keretrendszerbe
  - A prototípusunk a nagyobb biztonsági keretrendszer része, bővíteni kell más támadási kategóriákra is.
- Magyar nyelv támogatása
  - Magyar nyelvű adatbázis létrehozása.
- További modellek tesztelése
  - Alternatív, hatékonyabb modelleket keresése.



Köszönöm szépen a figyelmet!