DOI: https://doi.org/10.12700/btsz.2025.7.3.157

MANAGING INFORMATION SECURITY RISKS AND CORE VULNERABILITIES OF LARGE LANGUAGE MODELS

NAGY NYELVI MODELLEK ŐSSEBEZHETŐSÉGEINEK INFORMÁCIÓBIZTONSÁGI KOCKÁZATAI ÉS KEZELÉSÜK

BOTTYÁN Sándor[1]

Abstract

The spread of large language models (LLMs) has created new opportunities in information management and enterprise efficiency, while simultaneously exposing a several of new cybersecurity vulnerabilities. This study aims to map the most critical risks associated with LLMs, with particular emphasis on the protection of organizational information assets and the assurance of electronic information security. Based on a comparative analysis of the OWASP Top 10 for LLM publications, the research identifies "core vulnerabilities" that persistently feature within the constantly evolving threat landscape. The study analyzes the technical and organizational risks associated with these vulnerabilities, as well as their impact on the confidentiality, integrity, and availability of LLM ecosystems. Furthermore, it proposes the introduction of risk management and compliance mechanisms that holistically address the security challenges arising from LLMs at both organizational and technological levels.

Keywords

LLM vulnerabilities, prompt injection, data leakage, model poisoning, risk management

Absztrakt

A nagy nyelvi modellek (LLM-ek) elterjedése új lehetőségeket teremt az információkezelésben és a szervezeti hatékonyságnövelésben, azonban számos új típusú kiberbiztonsági sebezhetőséget is felszínre hozott. Jelen tanulmány célja, hogy feltérképezze az LLM-ek legkritikusabb kockázatait, különös tekintettel a szervezeti adatvagyon védelmére és az elektronikus információbiztonság biztosítására. A kutatás az OWASP Top 10 for LLM kiadványok összehasonlító elemzésén alapulva meghatározza azon „őssebezhetőségeket”, amelyek az állandóan változó fenyegetési környezet permanens szereplői. A tanulmány elemzi e sebezhetőségek technikai és szervezeti kockázatait, valamint azok hatásait az LLM-ökoszisztémák bizalmasságára, integritására és rendelkezésre állására. Ezen túlmenően javaslatot tesz olyan kockázatkezelési és megfelelőségi mechanizmusok bevezetésére, amelyek holisztikus módon, szervezeti és technológiai szinten egyaránt kezelik az LLM-ekből fakadó biztonsági kihívásokat.

Kulcsszavak

LLM sebezhetőségek, prompt injektálás, adatszivárgás, modellmérgezés, kockázatkezelés

[1] bottyan.sandor@gmail.com | ORCID: 0000-0002-9195-447X | PhD Student, University of Public Service Doctoral School of Military Engineering | doktorandusz, Nemzeti Közszolgálati Egyetem Katonai Műszaki Doktori Iskola munkakör magyarul, munkahely magyarul

BEVEZETÉS

Napjaink információs társadalmának legfőbb mozgatórugója az információ előállítása, terjesztése, valamint annak eredményes felhasználása. E fundamentum védelme, vagyis az információbiztonsági tevékenység kiemelt jelentőséggel bír. Az utóbbi évtizedek során tapasztalt digitalizáció eredményeként egyre nagyobb hangsúly helyeződik a kibertéri reziliencia kialakítására is, amelynek fontosságát a 2025. január 01-jén hatályba lépő, Magyarország kiberbiztonságáról szóló törvény megalkotásával a jogalkotó hangsúlyozta [1]. Ez az új szabályozó a NIS2 irányelvben meghatározott kiemelten kritikus és egyéb kritikus ágazatok szabályozásával párhuzamosan, részben azonos szervezet- és rendszerszintű védelmi intézkedéseket követel meg hazánk államigazgatási szerveitől és többségi állami tulajdonban lévő szervezetektől egyaránt [2]. Közismert tény, hogy a jog jellemzően a technológia után követője, hiszen folyamatosan jelennek meg olyan új, innovatív megoldások, amelyek túllépnek a már szabályozott kereteken. Ilyen jelenséget képviselnek napjainkban a nagy nyelvi modellek (large language models, a továbbiakban: LLM-ek) is, amelyekkel leginkább a népszerű csevegőrobotok (például: ChatGPT, Gemini, DeepSeek AI stb.) architektúrájában találkozhatunk és amelyek az utóbbi időszakban jelentős társadalmi népszerűségre és gazdasági térnyerésre tettek szert. Ez a technológia várhatóan rövid időn belül a szervezeti hatékonyság egyik alappillérévé válhat, hiszen az alkalmazásukban rejlő előnyök – például a fejlett információfeldolgozó képesség vagy a generatív funkciók – könnyen integrálhatók az információkezelési folyamatokba. Mindazonáltal az LLM-ek alkalmazása egyedi kockázatok formájában új kihívásokat jelent az elektronikus információbiztonság számára. Egyrészt annak sebezhetőségei által közvetlen veszélybe kerülhet a technológiát alkalmazó szervezet eddig kezelt adatvagyona, másrészt a saját tulajdonban működtetett nagy nyelvi modellek és rendszerkörnyezetük önmagában is jelentős értékű adatvagyon. Ennek okán az információvédelem szempontjából is kritikus jelentőségű az LLM-ekkel kapcsolatos biztonsági kérdések kezelése, ennek érdekében ez tanulmány röviden áttekinti a nagy nyelvi modellek legjelentősebb kiberbiztonsági sérülékenységeit, valamint javaslatokat fogalmaz meg ezek hatékony kezelésére.

A NAGY NYELVI MODELLEK TÉRNYERÉSE A SZERVEZETI TEVÉKENYSÉGEKBEN

A köznyelvben megfogalmazott mesterséges intelligencia népszerűségének egyik fő mozgatórugója a piacvezető chatbotok képességei. Ezek kifejezetten olyan emberi társalgásra finomhangolt (fine-tuned) LLM modellt és egyéb funkciókat is integráló architektúrák, amelyek ingyenesen bárki számára elérhetőek és természetes nyelveken, emberszerű vállaszokkal képesek valós időben kommunikálni a felhasználókkal. Mindemellett széleskörű tudással rendelkeznek a világunkról és képesek a kérésünknek megfelelő kreatív tartalmat generálni, ezáltal hasznosak a mindennapi tevékenységek (például: munkavégzés, tanulás, szórakozás) során. Ezt az előnyt minden versenyképességre törekvő szervezet, vagy állami feladatokat ellátó aparátus szívesen látná saját eszköztárában, így egyre elterjedtebb jelenség az LLM alkalmazása önálló szervezeti folyamatokban. A szoftvergyártók körében is egyre gyakoribb az asszisztencia jellegű felhasználása, több munkakörnyezeti szoftverben jelenik meg (pl. Microsoft Office 365 – Copilot, Google Workspace – Assistant). Kétségtelen, hogy ez a technológia hamarosan a szervezeti folyamatok technikai támogatójává

válik, ezáltal a szervezeti adatvagyonnal való jelentős kölcsönhatása elkerülhetetlen. Emiatt az elektronikus információbiztonságot érintő modell-sebezhetőségek, valamint a kapcsolódó kockázatok feltérképezése kifejezetten ajánlott tevékenység azon szereplők számára, amelyek eltökéltek az információbiztonsági politikában gyakorta foglalt alapelvek érvényesítésében.

A NAGY NYELVI MODELLEK RENDSZERSZINTŰ SEBEZHETŐSÉGEI

Az LLM-ek számottevő szervezeti szintű alkalmazása a ChatGPT (GPT-3.5) 2022 novemberi megjelenésével indult, hiszen ezt követően vált széles körben ismerté. Megkezdődött az interaktív webes felület jelentős használata, valamint az API eléréssel történő alkalmazása. A szervezeti alkalmazás ez esetben nem egyfajta szabályozott eljárásként, vagy saját fejlesztésként értelmezendő, hanem az úgynevezett „shadow AI" jelenség megjelenésére, amely az AI eszközök – informatikai vagy információbiztonsági szakterület jóváhagyása vagy felügyelete nélküli – alkalmazását jelenti a munkavállalók által. Az új technológia biztonsági tesztelésére számos önjelölt szakértő mellett kiberbiztonsági szervezetek is vállalkoztak. Az elmúlt években (2023-2024) az Open Web Application Security Project (OWASP) nemzetközi csapata több száz szakértővel és szervezettel (pl. AI óriásvállalatok, kiberbiztonsági vállatok, hardvergyártók, egyetemi kutatólaborok) együttműködve meghatározta a nagy nyelvi modellekkel működő alkalmazások legkritikusabbnak gondolt sebezhetőségeinek listáját. Tekintettel arra, hogy az LLM alkalmazások közege is folyamatosan változó környezet, vagyis úgynevezett „mooving landscape" (a kifejezés a kiberbiztonsági kockázatelemzésben arra utal, hogy az informatikai és technológiai környezet folyamatosan változik, ezért a kockázatok is állandóan átalakulnak), az eltérő verziójú listákban (0.1.0, 0.5.0, 0.9.0, 1.0.0, 1.0.1, 1.1.0 és 2025) rögzített LLM sebezhetőségek természetesen nem mutatnak homogén képet. Azonban egy mátrixban megjelenítve könnyen megállapíthatók azon típusú sebezhetőségek, amelyek az eltérő időszakokban történt felülvizsgálat alkalmával minden esetben megjelentek. Ezeket nevezhetjük úgymond „ősszebezhetőségeknek" is. (1. táblázat).

Kiadás éve	2023			2024
Verzió	0.1.0	0.5.0	0.9.0-1.1.0	2025
1.	Utasítás injektálás	Utasítás injektálás	Utasítás injektálás	Utasítás injektálás
2.	Adatszivárgás	Nem biztonságos kimenetkezelés	Nem biztonságos kimenetkezelés	Adatszivárgás
3.	Nem megfelelő sandbox izoláció	Tanítóadat-mérgezés	Tanítóadat-mérgezés	Ellátási lánc
4.	Jogosulatlan kódfuttatás	Szolgáltatás-megtagadás	Modell-szintű szolgáltatás-megtagadás	Adat- és modellmérgezés
5.	SSRF-sebezhetőségek	Ellátási lánc	Ellátási lánc	Helytelen kimenetkezelés
6.	Túlzott függőség	Jogosultsági problémák	Adatszivárgás	Túlzott önállóság
7.	Nem megfelelő MI-összehangolás	Adatszivárgás	Nem biztonságos bővítmények	Rendszerprompt-kiszivárgás
8.	Nem megfelelő hozzáférés-ellenőrzés	Túlzott önállóság	Túlzott önállóság	Vektor- és beágyazási gyengeségek

Kiadás éve	2023			2024
Verzió	0.1.0	0.5.0	0.9.0-1.1.0	2025
9.	Helytelen hibakezelés	Túlzott függőség	Túlzott függőség	Félrevezető információ
10.	Tanítóadat-mérgezés	Nem biztonságos bővítmények	Modell-lopás	Korlátlan erőforrás-felhasználás

1. Táblázat: a nagy nyelvi modellek legjelentősebb kiberbiztonsági sebezhetőségei. Forrás: a szerző szerkesztése

A fenti mátrixban rögzített sebezhetőségek mindegyike egyaránt kiemelten magas kockázatot képvisel, de a megjelenések gyakorisága alapján az alábbi három kritikus kockázatú össsebezhetőségeket határozhatjuk meg:

1. Prompt injektálás (prompt injection);
2. Adatszivárgás (data leakage or sensitive information disclosure);
3. Tanítóadat- és modellmérgezés (training data poisoning and model poisoning).

Ezek a sérülékenységek jellemzőik alapján kifejezetten differenciálhatók, továbbá az LLM architektúrák életciklusának több szakaszán is értelmezhetők. Az alábbiakban e sérülékenységek jellemzőinek részletes kifejtése következik.

Utasítás injektálás (prompt injection)

Az utasítás injektálás egyértelműen a nagy nyelvi modellek legkritikusabb sérülékenysége, amely során olyan speciálisan összeállított bemenetekkel (promptokkal) manipulálják a modellt, hogy annak működése kikerülje a rendszer alapvető biztonsági mechanizmusait, és a támadó által kívánt módon viselkedjen. A támadópromptok célja, hogy az architektúra valamely pontján megváltoztassa a modell működését. Két típusa ismert: a közvetlen és a közvetett. A közvetlen prompt injektálás során az LLM architektúra bemenetén (input) keresztül történik a sérülékenység-kihasználás. Ezzel szemben a közvetett prompt injection a modellnek egy későbbi működési folyamata során jut be az architektúra működésbe (például: plugineken vagy weboldalak forráskódján keresztül, beolvastatott fájl forráskódjában vagy szöveges tartalmában). [5] Szándék szerint ezek a támadások lehetnek szándékosak vagy szándékolatlanok. Előbbi esetében kifejezetten az sérülékenység kihasználásán érvényesül a rosszindulatú akarat, utóbbi esetében viszont valamilyen véletlen, előre nem várt esemény téríti el az alapértelmezett működést [6].

Ilyen típusú támadások során a támadók sikeresen kikerülhetik az LLM biztonsági mechanizmusait (safety layers), és legrosszabb esetben képesek felhasználni annak erőforrásait (például: számítási kapacitás, hálózati kapcsolatok), valamint különféle adatforrásokhoz (például: adatbázisok, konfigurációs beállítások stb.) való hozzáféréseit. Ez sajnos lehetőséget biztosít további rosszindulatú tevékenységek végrehajtására, például adatlopásra vagy social engineering típusú támadásokra is. Komplex támadási forgatókönyvekben egy kompromittált személyes LLM-asszisztens akár jelentős mennyiségű érzékeny információ kiszivárgását is eredményezheti, mivel hozzáfér a személyes levelezésünkben található információkhoz [7].

Kihasználása az alábbi főbb negatív hatásokat eredményezheti:

- Érzékeny információk kiszivárgása (például: tanítóanyag-, architektúra-, infrastruktúra információk; rendszer-promptok; személyes adatok; belső felhasználású és üzleti titkok stb.).
- Hibás vagy manipulált kimenetek létrehozása (például: döntéshozatali folyamatok manipulálása, genaratív funkciók érvényesülése hibás vagy manipulatív tartalommal, dezinformáció stb.).
- Jogosultság kiterjesztés (privilege escalation). A rosszindulatú személy a biztonsági funkciókat megkerülve többletjogosultságokat szerez az adott munkamenet (session) során, amellyel hozzáférhet a rendszer alapvető funkcióihoz és képessé válik tetszőleges parancsokat végrehajtattni, amely további kritikus hatásokat eredményezhet [6].

Adatszivárgás

Egy szervezeti adatvagyon integritásának és bizalmasságának megőrzése a nagy nyelvi modellek alkalmazása mellett tekintélyes információbiztonsági kihívás. Egyrészről az LLM generált kimenetén megjelenhetnek olyan szenzitív szervezeti adatok (például: személyes adatok, üzleti titkok), amelyek az előtanítás (pre-training), vagy a finomhangolási (fine-tuning) életciklus szakaszok során a tanítóanyagban – az adattisztítást követően is – megmaradtak. Másrészről a tanítóanyag – kiváltképp a webkaparással megalkotott korpuszok – tartalmazhatnak harmadik félre vonatkozó különleges adatokat (például: személyes adatok, jelszavak), amelyek jogtalan kezelésével és továbbításával adatvédelmi incidens jöhet létre [8]. Továbbá léteznek úgynevezett modell extrakciós támadások (model extraction attacks), amelyekkel meghatározható az LLM belső architektúrája és paraméterei, vagy akár tréningadatainak részleges rekonstrukciója is létrehozható. Amennyiben nem on-premise (olyan szoftvermegoldás, amely fizikailag kizárólag a szervezet saját infrastruktúráján található, nem pedig felhőalapú vagy külső szolgáltatónál), hanem úgynevezett hostolt modellek (például: ChatGPT, Gemini) képességét csatornázzuk be saját ügynök-alkalmazásunkba, ezen a kapcsolaton védett adatokhoz [9] férhet hozzá a harmadik fél (LLM üzemeltető). Ezeket az adatokat akár a modellek további fejlesztésére és tanítására is felhasználhatják [6].

Továbbá érdemes figyelembe venni az olyan humánkockázatokat, mint az architektúra jellemzőket tartalmazó rendszerterv dokumentációk szándékos kiszivárogtatása, vagy az ezt célzó social engineering támadás. Adatszivárgás jöhet létre továbbá a modellek hibás konfigurációs beállításaiból is, ez esetben „túlbeszélhet" a modell, vagyis a fejlesztési környezetben alkalmazott diagnosztikai vagy technikai információkat is megoszthat [6].

A sérülékenység kihasználása az alábbi főbb negatív hatásokat eredményezheti:

1. Személyes adatok (personal identifiable informations), illetve védett adatok kiszivárgása, amely a korábbi tanítóanyag rekonstruálásával, vagy az architektúrán keresztülhaladó adatfolyam (data stream) lecsapolásával történik.
2. Modellszivárgás (model leakage). A modell – akár szabadalom alatt álló –forráskódjainak, algoritmusainak, paramétereinek részbeni kiszivárgása, belső működésre vonatkozó információk kijutása [6].

Adatmérgezés vagy modellmérgezés

A modellek életciklusuk több pontján dolgoznak fel jelentős mértékű adatot, ezek az előtanítás (pre-traine), a finomhangolás (fine-tune), valamint az üzemelés alatti beágyazás (embedding). Adatmérgezésről beszélünk akkor, amikor a feldolgozandó adatkészletet rosszindulatú szándékkal úgy manipulálják, hogy az alapértelmezett modellműködés során konkrét sérülékenység, hátsó kapu (back door), vagy modell-elfogultság jöjjön létre. Egy sikeres adatmérgezés sajnos szignifikáns hatásokkal bír. Drasztikusan csökkenti a modell kiberbiztonsági állapotát és megítélését, ronthatja annak teljesítményét, eltérítheti a kívánt etikus viselkedési normáktól, közvetett hatásként számottevő kockázatot jelenthet a kapcsolódó ökoszisztémák biztonságára is. A legnépszerűbb, óriásvállatok által üzemeltett (hostolt) modellek kevésbé érzékenyek a modellmérgezésre a jelentős számban foganasított biztonsági intézkedésnek köszönhetően, míg a bárki számára elérhető nyílt forráskódú modellek (például: LLaMa, Mistral stb.) esetében valószínűbb mérgezési kockázatokkal kell számolnunk.

A sérülékenység kihasználása az alábbi főbb negatív hatásokat eredményezheti:

1. Eltérített modellműködés (1). A támadók előzetesen manipulálják a tanítóanyagot, amelyet később a modell képzéséhez a fejlesztők felhasználnak, ezt követően ez a generatív funkciók során befolyásolt kimeneteket eredményez.
2. Eltérített modellműködés (2). A támadók a rosszindulatú kódokat tartalmazó tartalmakat (például: weboldalak forráskódjai) állítanak elő (például: lejárt domainek újra vásárlása, Wikipedia oldal szerkesztése), amelyre az interakció során ráirányítják a modellt. A modell által beágyazott a káros kód eltéríti az alapértelmezett modellműködést.

KONKLÚZIÓ

A nagy nyelvi modellek gyors ütemű elterjedése és fejlődési üteme rendszeresen új típusú kiberbiztonsági sebezhetőségeket hoz felszínre, amelyek az LLM-ek életciklusának több pontján (például: tanítóanyag, előtanítás, finomhangolás, beágyazási folyamatok és interakció), is azonosíthatók. Hatásuk kiterjed a szervezeti adatvagyon bizalmasságára és sértetlenségére, a modelleket használó alkalmazások ökoszisztémájára és erőforrásaira, illetve a további kapcsolt rendszerekre egyaránt. Ebből adódóan az LLM ágensek (ügynökök) kiberbiztonsági kockázatkezelése összetett szakmai kihívás, hiszen egyaránt tartalmaz technológiai és szervezeti szintű intézkedéseket is. Előbbi csoportba olyan szükséges intézkedéseket sorolhatunk, mint az adattitkosítás, az adat-anonimizálás, a hozzáférés-ellenőrzés, az API (application programming interface) kezelés, az erős szintű autentikáció és autorizáció, valamint a sebezhetőség felderítés és a „secure by design” tevékenység. Míg az utóbbiba sorolható intézkedési kör többnyire a szervezetszintű információbiztonsági kockázatmenedzsment rendszer e tekintetben (pl. LLM sebezhetőség azonosítás, LLM incidensmenedzsment folyamatok való bővítését. Továbbá ide sorolható az AI compliance tevékenység, amely a GDPR, az AI Act követelményeknek való megfeleltetés mellett az etikai elvek alkalmazását, a folyamatok teljes átlátható dokumentálása is jelenti [3] [4]. A szervezetszintű biztonságos és felelős alkalmazásuk csak akkor biztosítható, ha a technológiai fejlődést a jogalkotás, a szakpolitika és a szervezeti gyakorlat egyaránt proaktívan követi és alakítja.Természetesen ez a tanulmány nem kívánja eme komplex rendszer minden elemét és

eljárását vizsgálni, megállapításaival leginkább a legfőbb LLM sebezhetőségekkel kapcsolatos kockázatkezelésen alapuló védekezéséhez kíván ajánlásokat megfogalmazni.

Ennek érdekében a tanulmány az alábbi fő megállapításokat tette:

- a különböző OWASP Top10 for LLM kiadások összehasonlításával megállapítható, hogy a LLM threat landscape dinamikusan változik, ugyanakkor az azonosított „őssebezhetőségek" tartós jelenléte kiemelt elemzési és kezelési fókuszt igényel,
- az egyes sérülékenységek konkrét, szervezeti szintű kockázati hatásai (például: jogosulatlan hozzáférés, adatlopás, működésmanipuláció stb.) azonosíthatók, hatásuk megbecsülhető. Ez lehetőséget ad a célzott védelmi és ellenőrzési mechanizmusok kialakítására.
- Az őssebezhetőségek szervezetszintű hatásai okán az LLM biztonságos alkalmazása csak holisztikus, szervezeti és technológiai szinten egyaránt összehangolt kockázatkezelési gyakorlat révén biztosítható.

JAVASLATOK

A hazánkban is érvényesített európai kiberbiztonsági, adatvédelmi vagy mesterséges intelligencia témában létrejött szabályozási irányelvek és rendeletek (például: NIS2, GDPR, AI Act) alapvető kiindulópontot nyújtanak ugyan, azonban nem reflektálnak teljes mértékben az LLM-ek sajátos működésére és életciklusából fakadó kiberbiztonsági kihívásaira. Az LLM-ek biztonsági és megfelelőségi követelményei a technológia működési sajátosságait is figyelembe vevő, szervezet és rendszerszintű szabályozási megközelítést igényelnek – különösen a magas kockázatú szektorok (pl. pénzügy, egészségügy, államigazgatás) esetében. Érdemes lehet a GRC (governance, risk management, and compliance) fogalomhoz köthető eljárásokat a nagy nyelvi modellek teljes életciklusát figyelembe véve megvalósítani. Szervezetszinten az AI irányítási rendszer (artificial Intelligence management system, AIMS), az adatvédelem, az információbiztonsági kockázatkezelés (különösen: information security managment system, ISMS), valamint a jogszabályi megfeleltetés folyamatait létrehozni. Míg rendszerszinten az architektúra-biztonság, az adatforrás-kontroll és a modellkimenetek validálhatósága válik a központi elemekké. Az ennek érdekében kialakított szervezeti eljárások létrehozása figyelembe vehetők az alábbi szabványok, ajánlások és jógyakorlatok:

- Az ISO/IEC 27000:2023 szabványcsalád alkalmazása. Különösen az ISO/IEC 27090 szabvány, amely útmutatást nyújt a szervezetek részére az mesterséges intelligencia rendszerek biztonsági fenyegetéseinek és hibáinak azonosítására és a kezelésére, és több más LLM sebezhetőség mellett az őssebezhetőségeket is kezeli [10]. Továbbá a kockázatok kezeléséhez járul hozzá az ISO/IEC 42000:2023 szabványcsalád is, amely e mellett támogatja a felelős és átlátható irányítást, valamint a jogszabályi megfelelést is.
- A NIST AI 600-1 szabvány, vagyis a mesterséges intelligencia kockázatmenedzsment keretrendszer (artificial risk management framework) az információbiztonság témáján belül szintén feldolgozza az tanulmány által megállapított főbb sebezhetőségeket. A szabvány alkalmazása az ISO/IEC 27000 alternatívájaként szintén segíthet a szervezeteknek azonosítani a generatív mesterséges intelligencia kockázatait,

és a szervezeti célokhoz illeszkedő kockázatkezelési mechanizmusok érvényre juttatását. [11]

- A Google Secure AI Framework (SAIF) biztonsági keretrendszerében foglalt kontrollok szintén alkalmazhatók az őssebezhetőségek kockázatainak kezelésére, a nagy nyelvi modellek biztonságos bevezetésére és működtetésére.
- Az LLM sebezhetőségekkel járó kockázatok csökkentésére, a sérülékenységek kihasználásának megelőzésére az OWASP ajánl konkrét stratégiákat, amelyek módszertanokat és technikai javaslatokat egyaránt tartalmaznak [6].
- A saját tervezésű LLM architektúrák megbiztonsági határainak megállapításához segítséget nyújthat az adatfolyam diagram (data flow diagram) alkalmazása, valamint a STRIDE (spoofing, tampering, repudiation, information disclosure, denial of service and elevation of privilege) módszertan is hozzájárulhat a sebezhetőségek azonosításához, ezátal a megfelelő védelmi intézkedések kidolgozásához [12].

FELHASZNÁLT IRODALOM

[1] 2024. évi LXIX. törvény Magyarország kiberbiztonságáról.
[2] Az Európai Parlament és a Tanács 2022. december 14-ei (eu) 2022/2555 irányelve az Unió egész területén egységesen magas szintű kiberbiztonságot biztosító intézkedésekről, valamint a 910/2014/EU rendelet és az (EU) 2018/1972 irányelv módosításáról.
[3] Az Európai Parlament és a Tanács 2016. április 27-i (EU) 2016/679 rendelete a természetes személyeknek a személyes adatok kezelése tekintetében történő védelméről és az ilyen adatok szabad áramlásáról, valamint a 95/46/EK irányelv hatályon kívül helyezéséről (általános adatvédelmi rendelet).
[4] Az Európai Parlament és a Tanács (EU) 2024/1689 rendelete (2024. június 13.) a mesterséges intelligenciára vonatkozó harmonizált szabályok megállapításáról, valamint a 300/2008/EK, a 167/2013/EU, a 168/2013/EU, az (EU) 2018/858, az (EU) 2018/1139 és az (EU) 2019/2144 rendelet, továbbá a 2014/90/EU, az (EU) 2016/797 és az (EU) 2020/1828 irányelv módosításáról (a mesterséges intelligenciáról szóló rendelet) (EGT-vonatkozású szöveg)
[5] „Wunderwuzzi's blog. Plugin Vulnerabilities: Visit a Website and Have Your Source Code Stolen," [Online]. Available: https://embracethered.com/blog/posts/2023/chat-gpt-plugin-vulns-chat-with-code/.
[6] „OWASP Top 10 for Large Language Model Applications.," [Online]. Available: https://owasp.org/www-project-top-10-for-large-language-model-applications/.
[7] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz és M. Fritz, Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection., https://doi.org/10.48550/arXiv.2302.12173, 2023.
[8] H. Nihad, „Explore mitigation strategies for 10 LLM vulnerabilities.," 2024. [Online]. Available: https://www.techtarget.com/searchEnterpriseAI/tip/Explore-mitigation-strategies-for-LLMvulnerabilities..
[9] 2023. évi CI. törvény a nemzeti adatvagyon hasznosításának rendszeréről és az egyes szolgáltatásokról.

[10] „ISO/IEC DIS 27090(en). Cybersecurity – Artificial Intelligence – Guidance for addressingsecurity threats and failures in artificial intelligence system," [Online]. Available: https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27090:dis:ed-1:v1:en.
[11] „NIST InformationTechnology Laboratory. AI Risk Managment Framework.," [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework.
[12] G. Klondik, „Threat Modeling LLM Applications.," 2023. [Online]. Available: https://aivillage.org/large%20language%20models/threat-modeling-llm/.
[13] „NIST Trustworthy and Responsible AI NIST AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence," [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.