DOI: https://doi.org/10.12700/btsz.2025.7.4.177

COMPARATIVE ANALYSIS OF LLM ARCHITECTURES IN SAFETY-CRITICAL SYSTEMS: DEPLOYMENT MODELS, SOVEREIGNTY, AND VULNERABILITY

LLM ARCHITEKTÚRÁK ÖSSZEHASONLÍTÓ ELEMZÉSE BIZTONSÁGKRITIKUS RENDSZEREKBEN: TELEPÍTÉSI MODELLEK, SZUVERENITÁS ÉS SEBEZHETŐSÉG

BEDNÁRIK Boldizsár[1] – GOGOLÁK László[2]

Abstract

This study examines the possibilities and risks of applying large language models (LLM) in safety-critical systems and demonstrates how LLMs can be integrated into decision making process (support) and automated systems. It provides a detailed comparison of on-premise, cloud-based, and hybrid deployment models, with particular focus on data sovereignty, information security, and operational flexibility. Based on OWASP and DHS guidelines, the paper outlines key vulnerabilities. It also analyzes the relevance of the NIS2 Directive, and the AI Act framework. To reduce risks, it highlights the importance of the ALARP principle, human oversight, validation mechanisms, multi-layered protection, and security-conscious training. The study's aim is to support a sustainable balance between safety and innovation.

Keywords

LLM, Data Sovereignty, Risk Management, Artificial Intelligence (AI), Data Security

Absztrakt

Ez a tanulmány a nagy nyelvi modellek (LLM) biztonságkritikus rendszerekben történő alkalmazásának lehetőségeit és kockázatait elemzi, valamint bemutatja, hogyan épülhetnek be az LLM-ek a döntéstámogató és automatizált rendszerekbe. Részletesen összehasonlítja a lokális, felhőalapú és hibrid telepítési modelleket, különös tekintettel az adatszuverenitásra, adatbiztonságra és működési rugalmasságra. Az OWASP és a DHS irányelvei alapján bemutatásra kerülnek a legfontosabb sebezhetőségek.A tanulmány elemzi a NIS2 irányelv és az AI Act előírásainak relevanciáját. A kockázatok csökkentésére az ALARP-elv, valamint emberi felügyelet, validáció, több rétegű védelem és biztonságtudatos oktatás kerül kiemelésre. A tanulmány célja a biztonság és innováció közötti fenntartható egyensúly elősegítése.

Kulcsszavak

LLM, adatszuverenitás, kockázatkezelés, Mesterséges Intelligencia (MI), adatbiztonság

[1] bednarik.boldizsar@uni-obuda.hu | ORCID: 0009-0005-5198-1210 | PhD Student, Doctoral School on Safety and Security Sciences, Óbuda University| Doktorandusz, Biztonságtudományi Doktori Iskola, Óbudai Egyetem

[2] gogolak@mk.u-szeged.hu |ORCID: 0000-0002-6653-3534 | Associate Professor, Department of Mechatronics and Automation, Faculty of Engineering, University of Szeged| Főiskolai Docens, Mechatronikai és Automatizálási Tanszék, Mérnöki Kar, Szegedi Tudományegyetem

BEVEZETÉS

A nagy nyelvi modellek (LLM-ek) rohamos fejlődése - mint megannyi más területen - új távlatokat nyitott a biztonságkritikus rendszerekben is - az energetikától és közlekedéstől a védelmi szférán át az egészségügyig. E rendszerekben az LLM-ek bevezetése egyszerre kecsegtet forradalmi előnyökkel és vet fel komoly biztonsági és etikai kihívásokat.

Az LLM-ek alkalmazása ezen a területen azért vált központi kérdéssé, mert segítségükkel nagy adatmennyiségek gyorsabban és könnyebben elemezhetők, ezzel lehetővé válhat az esetleges problémák előrejelzése, és az egyszerűbb - megközelítőleg valósidejű - automatizált döntéshozatal. Egy másik ok, hogy az LLM-ek képesek tanulni és összefüggéseket felfedezni olyan összetett rendszerekben és adathalmazokban, ahol az emberi szakértők kapacitása véges. Mindez különösen vonzó az olyan kritikus infrastruktúráknál, ahol az üzemzavar vagy támadás emberéleteket veszélyeztethet, jelentős anyagi-, személyi kárt okozhat, vagy nemzetbiztonsági kockázattal jár.

Ugyanakkor az LLM-ek integrálása biztonságkritikus környezetben nem pusztán technikai kérdés. A biztonságtudomány maga is multidiszciplináris terület: magában foglalja a műszaki, társadalomtudományi, jogi és etikai dimenziókat egyaránt. A biztonság fogalma komplex és multidimenzionális, több tudományág metszéspontjában áll, ami megnehezíti az egységes megközelítést [1]. Ez a mi esetünkben azt jelenti, hogy amikor LLM-alapú mesterséges intelligenciát építünk be például egy atomerőmű vezérlőrendszerébe vagy egy autonóm jármű irányításába, akkor nem elég csupán a gépi tanulás technikai teljesítményét néznünk. Figyelembe kell venni a humán tényezőt (pl. a kezelők bizalmát a rendszer iránt), a jogszabályi megfelelést (pl. adatvédelmi előírásokat), az etikai normákat (pl. dönthet-e gép emberi életet érintő kérdésben), és nem utolsó sorban a szervezeti és nemzeti szuverenitás kérdését is.

Az LLM-ek biztonságkritikus területen történő alkalmazása így valódi multidiszciplináris kihívás, amely érinti a különböző tudományterületek szempontjait. Ezt szemelőtt tartva tanulmányunk célja, hogy megpróbáljon átfogó képet adni az LLM-ek biztonságkritikus rendszerekben való alkalmazásáról, különös tekintettel a lehetséges telepítési modellekre és feltárja az ezzel járó dilemmákat.

TELEPÍTÉSI MODELLEK ÖSSZEHASONLÍTÁSA

Az LLM-ek integrációja során alapvető kérdés, hogy milyen telepítési modellt (deployment) válasszunk: házon belüli (on-premise), felhőalapú (cloud) vagy hibrid megoldást. Ezek a modellek jelentősen eltérnek az adatkezelés, hozzáférés, szuverenitás és biztonság szempontjából, ezért fontos a tudatos mérlegelés.

Az **On-premise (helyszíni) telepítés** alkalmazása esetében az LLM rendszert és az adatait teljes mértékben a saját, helyi infrastruktúrán futtatjuk. Ennek nagy előnye a teljes felügyelet az adataink felett: az érzékeny adatokat nem kell kiadni külső szolgáltatónak, így jobban biztosítható a megfelelés a szigorú iparági vagy nemzeti előírásoknak. Katonai vagy kormányzati alkalmazásoknál például elsődleges szempont, hogy az adat a rendszer határain belül maradjon. A NATO NCIA Technology Strategy 2030 is kiemeli, hogy a NATO műveleti környezetében a nyilvános, privát, hibrid, szuverén és edge felhőmegoldások egyaránt mérlegelendők a különböző biztonsági besorolású adatok és

szolgáltatások esetén [2]. Az on-premise LLM tehát maximális kontrollt nyújt, ugyanakkor hátránya a magas iniciális költség és erőforrásigény: a nagy modell futtatásához szükséges hardver-infrastruktúrát ki kell építeni és karban kell tartani, a modell fejlesztését és frissítését saját erőből kell megoldani. Ez pénzben és időben is jelentős terhet róhat a szervezetre. Ráadásul skálázhatósági korlátok is lehetnek: a felhasználói igények gyors növekedése esetén nehézkes lehet lépést tartani a szükséges szerverkapacitás bővítésével.

Az LLM **Cloud (felhőalapú) telepítésénél** a szolgáltatást egy felhőszolgáltató infrastruktúráját használva vesszük igénybe (pl. nagy tech cégek AI platformjain). Előnye a gyors fejlesztési és bevezetési ciklus: a felhőben általában azonnal rendelkezésre áll a szükséges számítási kapacitás, valamint hozzáférhetők a legkorszerűbb pre-trainelt nyelvi modellek. Ezáltal a fejlesztés felgyorsul, és a szolgáltatás rugalmasan skálázható (nagy terhelés esetén automatikusan több erőforrást kapunk). A felhőszolgáltatók gyakran erős beépített biztonsági mechanizmusokat kínálnak, folyamatos frissítésekkel, és nagyfokú redundanciával üzemeltetik rendszereiket. Ugyanakkor a felhőmodell jelentős adatvédelmi kockázatot hordoz: az adatok kikerülnek a saját infrastruktúránkból, ami aggályos lehet pl. személyes vagy érzékeny adatok esetén. A felhőszolgáltatóval szemben bizalomra van szükség - nem mindegy, hogy a cég mely ország joghatósága alá tartozik, és hogyan garantálja az adatszuverenitást. Gyakori elvárás, hogy szuverén felhő opció álljon rendelkezésre, ahol biztosított az adatok országon vagy szövetségi kereteken belüli tárolása és kezelése [3]. A NATO pl. az Oracle által kínált szuverén felhőmegoldásokra való áttérés célja, hogy a legújabb felhő- és AI-innovációkat a kritikus adatok biztonságának feláldozása nélkül vehessék igénybe [3]. Meg kell még említeni a felhőmodellel kapcsolatos másik kockázatot is, ami a szolgáltatótól való függés: a vendor lock-in és a szolgáltatás kimaradása (downtime) is veszélyeztetheti a működést. Továbbá egy nyilvános felhőben futó LLM esetén számolni kell azzal is, hogy több-bérlős környezetben fut a rendszer, potenciálisan megosztva erőforrásokat ismeretlen ügyfelekkel, ami kifinomult támadások (pl. oldalcsatornás támadások) kockázatát felvetheti [4].

A **Hibrid** megközelítés igyekszik ötvözni az on-premise és felhő előnyeit. A kritikus vagy érzékeny adatokat és funkciókat házon belül tartja, míg a kevésbé érzékeny feladatokat, valamint a csúcsterhelés esetén szükséges többletkapacitást a felhőbe delegálja. Például egy egészségügyi alkalmazásnál a páciensek személyes adatait helyben tárolhatják és kezelhetik, míg az általános nyelvi elemző modellt felhőben futtatják. A NATO stratégiában is megjelenik a multi-cloud integráció igénye, melynek célja, hogy egységes szolgáltatásként lehessen kezelni a különböző (privát és publikus) felhők erőforrásait [2].

Összességében a telepítési modell megválasztása szorosan összefügg a szuverenitás kérdésével. Biztonságkritikus rendszerekben sokszor a nemzeti/üzleti szuverenitás megőrzése az elsődleges szempont, ami az on-premise vagy szuverén felhős megoldások felé billenti a mérleget. Azonban, ahol gyors innovációra és skálázhatóságra van szükség, ott a felhő előnyei (sebesség, rugalmasság, költséghatékonyság) előtérbe kerülnek, persze csak abban az esetben, ha megfelelő szerződéses és technikai garanciákkal kezelik az adatvédelmi kockázatokat. A következőkben áttekintjük az ilyen rendszerekben rejlő főbb előnyöket és kockázatokat.

ELŐNYÖK ÉS KOCKÁZATOK

Egy LLM alkalmazása biztonságkritikus rendszerben számos előnyt kínál: képes komplex mintázatokat felismerni az adatokban, szabályalapú rendszerekkel nem detektálható anomáliákra figyelmeztetni, előre jelezni várható eseményeket (pl. berendezéshibát vagy kibertámadást), sőt bizonyos mértékű autonóm döntést is hozhat rutinszerű helyzetekben [5]. Ezáltal növelheti a rendszer hatékonyságát (gyorsabb reakcióidő, 24/7 rendelkezésre álló "mesterséges szakértő"), és tehermentesítheti az emberi operátorokat a monoton vagy túl sok adatot igénylő feladatok alól. Például egy energiatermelő hálózatban az LLM képes lehet a szenzoradatokból és üzemzavari jegyzőkönyvekből tanulva előre jelezni egy transzformátor meghibásodását, megelőzve ezzel a nagyobb balesetet vagy kiesést [6]. A kórházi környezetben egy LLM segíthet gyorsan áttekinteni az orvosi protokollokat és javaslatot tenni a sürgősségi ellátás priorizálására, amikor nagyszámú sérült érkezik. Ezek az előnyök - a gyorsaság, skálázhatóság, tudás-integráció - teszik vonzóvá az LLM-eket a kritikus alkalmazásokban.

De beszélni kell a másik oldalról is: az LLM-ek bevezetése új kockázatokat és sebezhetőségeket hoz a rendszerbe, amelyeket nem szabad alábecsülni. A DHS (Amerikai Belbiztonsági Minisztérium) 2024-es AI-CI iránymutatásai három fő kockázati kategóriát különböztetnek meg a kritikus infrastruktúrák kapcsán [7]:

1. Az AI által végrehajtott támadások - amikor a támadók magukat az LLM-eket használják fel fegyverként, például a kibertámadások automatizálására vagy a félrevezető álhírek tömeges generálására
2. Az AI rendszerek elleni támadások - amikor magát az LLM-alapú rendszert veszik célba, például adverszáriális (megtévesztő) módszerekkel manipulálják a bemeneteket (evasion támadások) vagy adatmérgezéssel (data poisoning) rontják le a modell teljesítményét [7] [8]
3. Az AI tervezési hibáiból fakadó problémák - ide tartoznak a rendszer belső gyengeségei, úgymint a túlzott autonómia, a törékenység (brittleness - amikor váratlan bemenetre kiszámíthatatlanul reagál a modell) vagy az értelmezhetetlenség, ami miatt az AI döntéseit nem lehet megmagyarázni [7]

Ezek a kategóriák rávilágítanak, hogy az LLM-ek kapcsán nem csak a külső rosszindulatú szereplőktől kell tartanunk, hanem a rendszer belső hiányosságaitól is.

Fontos még az etikai és jogi kockázatok figyelembevétele. Egy LLM által hozott döntés diszkriminatív vagy igazságtalan lehet, ha a modell torz adatokon tanult - például az egészségügyben előfordulhat, hogy egy AI rendszer alulértékeli egyes kisebbségi csoportok rizikófaktorait a tanulási adatok hiányosságai miatt. A felelősség kérdése is felmerül: ki a felelős, ha egy LLM hibás javaslata kárt okoz? Jogi szempontból az EU friss AI Act (2024/1689) rendelete kockázati szintek alapján szabályozza az AI rendszereket, és a magas kockázatú alkalmazásokra (ahová a biztonságkritikus felhasználások is tartoznak) szigorú átláthatósági, biztonsági és megfelelőségi követelményeket ír elő [9]. Hasonlóképpen, az EU NIS2 irányelv (Directive (EU) 2022/2555) kibővíti a létfontosságú ágazatok kibervédelmi előírásait, kötelező kockázatkezelési gyakorlatokat és incidens-jelentési rendszert vezet be - ez közvetett módon az AI-alapú komponensekre is vonatkozik, hiszen azokat is integrálni kell a teljes kockázatmenedzsment folyamataiba [10]. Végül, a szuverenitási aggályok is ide tartoznak: például, ha egy kritikus infrastruktúra külső,

külföldi AI szolgáltatást használ, az adott ország kiszolgáltatottá válhat egy idegen entitásnak, ami geopolitikai kockázatot hordozhat.

Összefoglalva, az LLM-ek alkalmazása a biztonságkritikus rendszerekben egyszerre ígéretes és kockázatos: a mérleg egyik serpenyőjében az intelligens automatizáció és hatékonyságnövelés áll, a másikban pedig az új támadási felületek, tervezési hibák és szabályozási dilemmák. A következő szakaszban egy összehasonlító mátrix segítségével szemléltetjük, hogy különböző alkalmazási területeken hogyan viszonyul egymáshoz a várható előny és a kockázati szint.

ALKALMAZÁSI TERÜLETEK ÖSSZEHASONLÍTÁSA A FELMERÜLŐ KOCKÁZAT ÉS NYÚJTOTT ELŐNYÖK SZEMPONTJÁBÓL

Az 1. táblázat hat lehetséges alkalmazási területen szemlélteti az LLM-ek biztonságkritikus használatának kockázati szintjét (alacsony / közepes / magas) és a várható előnyt (alacsony / közepes / magas). Minden cellában rövid értékelés foglalja össze a kockázat-előny mérlegét, indokolva a besorolást. Az értékelések általános, útmutatás jellegűek. Jól látszik az AI alkalmazás biztonság/előny kettős természete.

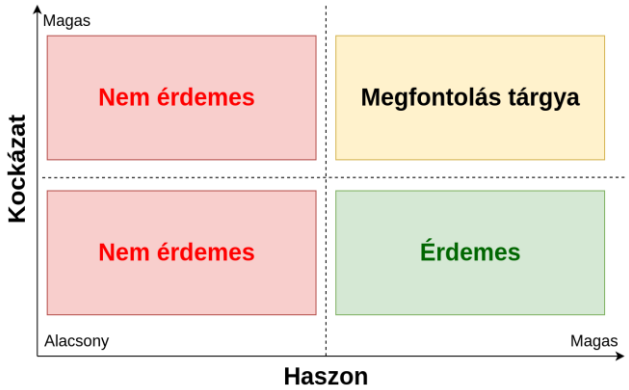
Alkalmazási terület	Fő előny	Fő kockázat	Kockázati szint	Potenciális előny
Kritikus infrastruktúra üzemeltetés	**Gyors hibadetektálás, prediktív karbantartás**	**Automatizált döntések hibás működése, modellmanipuláció**	Magas	Közepes-magas
Egészségügyi döntéstámogatás	**Diagnosztikai pontosság növelése, adminisztrációs teher csökkentése**	**Életet veszélyeztető hibás javaslat, adatvédelmi incidens**	Nagyon magas	Magas
Közigazgatás és e-kormányzat	**Ügyintézés gyorsítása, emberi erőforrás tehermentesítése**	**Adatszivárgás, torzított döntéstámogatás**	Közepes	Magas
Oktatás és tudásmenedzsment	**Személyre szabott tanulás, tartalomgenerálás**	**Félrevezető információ, túlzott automatizálás**	Alacsony-közepes	Közepes
Pénzügyi szektor / FinTech	**Csalásdetektálás, automatizált kockázatelemzés**	**Adatszivárgás, modell-torzítás, megfelelőségi kockázat**	Magas	Magas

Vállalati támogatórendszerek (HR, marketing, ügyfélszolgálat)	**Produktivitásnövelés, gyors információáramlás**	**Bizalmas adatok kiszivárgása, torz eredmények**	Alacsony-közepes	Magas

1. táblázat: Egyes alkalmazási területek becsült haszna és kockázata a különböző telepítési modellek (on-premise, cloud, hibrid) függvényében

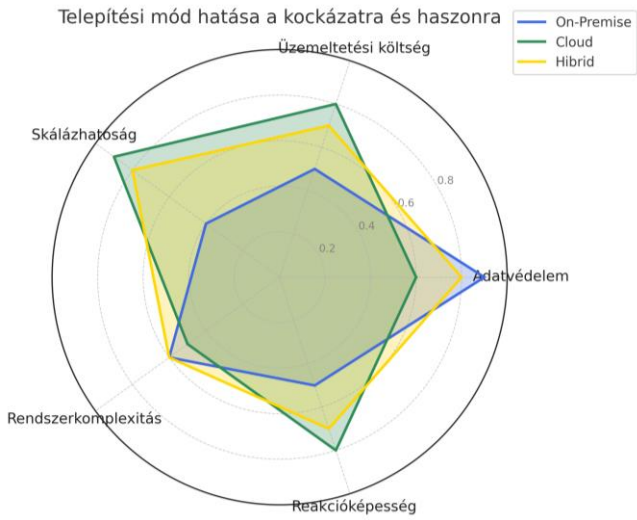
A döntéshozatal egyik fő eleme a kockázat - haszon egyensúly megfelelő értékelése. Az ALARP (As Low As Reasonably Practicable) elv szerint a kockázatot addig kell csökkenteni, amíg az ésszerűen megvalósítható - azaz amíg a kockázatcsökkentés költségei és erőforrásigénye arányban állnak a várható haszonnal [11].

A 1. ábra szemlélteti hogy mely zónákban tekinthető az alkalmazás indokoltnak, útmutatóként a döntéshozatalhoz.



1. ábra: A kockázat és a haszon viszonya négy döntési kategóriában.

A telepítési modell kiválasztása döntő szerepet játszik abban, hogy az AI-rendszer milyen kockázat-haszon arányt képvisel. A cloud környezetek magasabb hasznot kínálnak a skálázhatóság és az innováció gyors elérhetősége révén, ugyanakkor növelik az adatvédelmi és szolgáltató-függőségi kockázatokat. Az on-premise rendszerek ezzel szemben fokozott biztonságot és kontrollt adnak, de nagyobb beruházási és üzemeltetési költségekkel járnak. A hibrid megközelítés az ALARP-elvnek megfelelő kompromisszumot jelentheti, ahol a kockázat és a haszon egyensúlyban tartható. A 2. ábra szemlélteti ezeket az összefüggéseket.



2. ábra: A különböző telepítési modellek eltérő mértékben befolyásolják a mesterséges intelligencia rendszerek kockázati és haszon-profilját.

AJÁNLÁSOK, MEGELŐZÉS ÉS KÁRENYHÍTÉS

A fenti elemzés alapján kirajzolódik, hogy az LLM-ek biztonságkritikus alkalmazása során az egyik kulcskérdés a kockázatok megfelelő csökkentése. Az ALARP-elv (As Low As Reasonably Practicable - amilyen alacsony szintre csak ésszerűen csökkenthető) alkalmazása megkerülhetetlen: törekedni kell arra, hogy minden azonosított kockázatot olyan alacsony szintre mérsékeljünk, amennyire az ésszerűen megvalósítható [11]. Ahogyan a biztonságtechnikában általános, el kell fogadnunk, hogy zéró kockázat nem létezik (fokozottan igaz ez a komplex AI-rendszerekre), de a fennmaradó kockázat soha nem haladhatja meg az elfogadható szintet. Ha egy további védelmi intézkedés már csak aránytalanul nagy ráfordítással lenne bevezethető a várható nyereséghez képest, akkor teljesül az ALARP kritérium [11]. Ezt az elvet követve az LLM-ek esetében is minden ésszerű biztonsági kontrollt implementálni kell, és csak azokról mondjunk le, amelyek költsége vagy komplexitása már nincs arányban a hozott biztonsági előnnyel. Tekintsünk át néhány ajánlást, bevált kárenyhítési javaslatot és gyakorlatot:

Többrétegű védelem alkalmazása

Az LLM-ek köré többszintű védelmi mechanizmusokat kell építeni (defense-in-depth). Egyetlen intézkedés sem véd minden ellen, de több, egymást kiegészítő kontroll kombinációja erős biztonsági hálót ad. Például: a rendszer inputját egy előszűrő modul ellenőrizze, mielőtt az LLM-hez jut (kiszűrve a nyilvánvalóan káros vagy injekciós próbálkozásokat); az LLM kimeneteit egy utóellenőrző modul vizsgálja, mielőtt a külvilág felé továbbléptetnénk (pl. ne generálhasson közvetlenül parancsot egy vezérlőrendszernek emberi jóváhagyás nélkül). Emellett gondoskodni kell a hálózati szintű védelemről (tűzfalak, behatolásészlelés), és az LLM-et futtató környezet izolációjáról (konténerezés, homokozó jellegű futtatás), hogy egy esetleges kompromittálás ne terjedhessen tovább a rendszerben.

Bemeneti és kimeneti validáció (valódiság-ellenőrzés)

Az LLM-ek sajátossága, hogy bármilyen utasítást követnek a promptban, hacsak nincsenek korlátok közé szorítva. Az OWASP útmutatók is hangsúlyozzák: a modellt kezeljük úgy, mintha egy külső felhasználó lenne - ne bízzunk vakon a generált válaszában [12]. Implementáljunk szigorú bemeneti (input) validációt: engedélyezett parancsok és formátumok listáját (allowlist) használva, és utasítsuk el vagy semlegesítsük (és naplózzuk!) a gyanús bemeneteket. Például egy LLM ne kaphasson olyan rendszerpromptot a felhasználótól, ami érzékeny információ kiadására vagy tiltott műveletre buzdítja. A kimeneteknél pedig alkalmazzunk kontextusfüggő kódolást és szűrést: például ha az LLM kódot vagy parancsot generál (pl. SQL lekérdezést, shell parancsot), azt csak megfelelő escaping/encoding után használjuk fel [12]. Egy webes megjelenítésnél pl. HTML kimenetet escape-elni kell, hogy megelőzzük az XSS támadást, még akkor is, ha azt az LLM "magától" írta. Soha ne futtassuk egy LLM kimenetét közvetlenül a rendszerben ellenőrzés nélkül (ne hívjunk meg eval-t a generált kódra, ne küldjük el nyers formában e-mailben stb.). Az ilyen validációs-szűrési lépések automatizálhatók és beépíthetők az alkalmazás logikájába.

Prompt injection elleni védelem

A prompt injection - amikor a támadó olyan bemenetet ad, amivel átveszi az irányítást a modell kimenete felett - az LLM-ek speciális Achilles-sarka. Ennek kivédésére kombinált módszerek szükségesek. Egyrészt használjunk statikus prompt sablonokat, amelyekből a felhasználó által beírható részt minimálisra szorítjuk, és nem engedjük felülírni a rendszer üzeneteket. Másrészt monitorozzuk a párbeszédeket automatizáltan: gyanús minták detektálására (pl. ha a felhasználó arra utasítja az LLM-et, hogy ne tartsa be a korábbi utasításokat, vagy adjon ki tiltott tartalmat). Megfelelő eszközökkel felismerhetők az ilyen injekciós kísérletek, és ilyenkor a rendszer adjon egy generikus választ vagy tagadja meg a kérést. Továbbá a figyelmeztető jelzések (pl. a modell hirtelen a rendszerprompt érzékeny részleteit akarja felfedni) indítsanak el egy incidenskezelési folyamatot. Nem utolsó sorban pedig tartsuk naprakészen a modellünket: az újabb LLM verziók tipikusan javítanak a biztonsági korlátokon, az ismert prompt injection trükkök ellen védettebbek. Amennyiben saját modellt használunk, rendszeresen finomhangolhatjuk ellenpéldákkal (adversarial training), hogy ellenállóbb legyen az ilyen támadási mintákra.

Emberi felügyelet és vészfunkciók

Bármilyen fejlett is legyen egy AI, biztonságkritikus rendszerekben nem szabad felügyelet nélkül hagyni. Elengedhetetlen egy olyan emberi felügyeleti mechanizmus, amely folyamatosan követi az LLM működését, és szükség esetén közbe tud lépni. Ez jelentheti azt, hogy az LLM döntései vagy ajánlásai csak emberi jóváhagyással válnak végrehajthatóvá (human-in-the-loop), vagy legalább azt, hogy egy operátor vissza tud vonni/ki tud kapcsolni egy AI által indított műveletet (pl. ha az AI tévesen egy rossz lépést javasolna, az ember felülbírálhatja). Különösen az autonómabb funkciók esetén legyen egy "vörös gomb" (kill switch), amivel az AI modul gyorsan leállítható, ha rendellenes működést produkál. A felügyelet része a monitoring: valós idejű naplózás és riasztások arról, hogy az LLM milyen döntéseket hoz, milyen bizalmas adatot kér vagy ad ki.

Biztonságtudatos tervezés és oktatás

Az LLM-ek biztonságos használata nem csak technikai kérdés, hanem emberi tényező is. Gondoskodni kell róla, hogy mindazok, akik a rendszert tervezik, üzemeltetik

vagy használják, megfelelő képzést kapjanak az AI sajátosságairól és kockázatairól. A fejlesztők számára szervezzünk biztonsági tréninget kifejezetten AI témában (pl. OWASP LLM Top 10 ismertetése, tipikus támadási vektorok bemutatása) [12]. Az üzemeltető csapat ismerje az LLM monitorozási eszközeit, tudja értelmezni a riasztásokat és legyen kész vészhelyzeti eljárásrend (pl. ha azt gyanítják, hogy az AI kompromittálódott, hogyan vegyék ki a rendszerből). A végfelhasználóknak - pl. egy erőmű vezérlőtermében dolgozó mérnököknek vagy kórházi orvosoknak - pedig oktatni kell, hogy mire képes és mire nem képes a bevezetett LLM-alapú asszisztens. Fontos, hogy megértsék: az AI által adott tanács nem szentírás, kritikusan kell kezelni. Tanítsuk meg nekik az alapvető „AI-higiéniát": ne adjanak be a rendszernek értelmetlen vagy rosszindulatú utasítást (hiszen ezzel ők maguk is előidézhetnek hibás működést), figyeljenek a rendszer szokatlan válaszaira, és jelentsék, ha valami gyanúsat tapasztalnak. A szervezeti biztonsági kultúra kiterjesztése az AI területére kulcsfontosságú - a biztonságtudatosság növelése az egyik leghatékonyabb, bár gyakran alulértékelt védelmi vonal [13].

Ezen ajánlások egy része már megjelent a nemzetközi szabványokban és ajánlásokban, más részüket a józan ész és tapasztalat diktálja. A lényeg az, hogy holisztikusan, a szervezet egészére kiterjedően kezeljük az LLM-ekből fakadó biztonsági kihívásokat - egyszerre fókuszálva a technológiára (secure coding, monitoring, hardening) és a szervezetre (irányítás, folyamatok, képzés) [13]. Így minimalizálhatjuk annak esélyét, hogy egy AI miatt következzen be súlyos incidens, illetve, ha mégis bekövetkezik, felkészülten, jól begyakorolt tervvel reagálhassunk.

KORLÁTOK ÉS TOVÁBBI KUTATÁS

A tanulmány elsősorban az LLM-ek biztonságkritikus környezetekben való alkalmazásának elméleti és tervezési szempontjaira összpontosít, tehát a bemutatott megközelítés elméleti keretet ad, de működésének gyakorlati ellenőrzése még nem történt meg, és nem tartalmaz konkrét (mennyiségi) kockázati mutatókat sem. A telepítési modellek összehasonlítása is jelenleg leíró jellegű, ahogy a költség-haszon elemzés és a teljesítménymérés is további kutatási feladat lesz.

KÖVETKEZTETÉS

A nagy nyelvi modellek integrációja a biztonságkritikus rendszerekbe egyaránt hordoz magában óriási lehetőségeket és jelentős felelősséget. Az LLM-ek lehetőséget nyújtanak arra, hogy a biztonságtechnika területén új szintre lépjünk: intelligensebb felügyeleti rendszerek, proaktív hibamegelőzés, gyorsabb és pontosabb vészhelyzeti reagálás valósulhat meg általuk. Például képzeljünk el egy okos elektromos hálózatot, ahol az AI előre jelzi és megakadályozza a hálózati túlterhelést, vagy egy önvezető jármű-flottát, mely minimális emberi beavatkozással, de maximális biztonsággal közlekedik. Ezek nem távoli sci-fi víziók, hanem reális középtávú célok, ha az LLM-eket sikerül megfelelő korlátok között, biztonságosan alkalmazni, ehhez azonban elengedhetetlen, hogy ezek a rendszerek alkalmazhatóságát mindig a társadalom és az üzemeltetők bizalom-küszöbéhez mérjük. Ahogy a biztonságtudományban alapelv, itt is csak akkor tekinthető elfogadhatónak egy új technológia bevezetése, ha a hozzá kapcsolódó kockázatok az ALARP-elv szerint kellően lecsökkentek. Ez a gyakorlatban azt jelenti, hogy addig nem szabad egy LLM-et

éles kritikus környezetbe állítani, amíg minden ésszerű biztonsági intézkedést meg nem tettünk, és a megmaradt, tovább nem csökkenthető kockázatot el tudjuk fogadni.

A nagy nyelvi modellek a biztonságtechnika fejlődésének következő mérföldkövét jelenthetik, de csak akkor, ha bevezetésük okosan, felelősen történik. A jövő biztonságkritikus rendszerei talán már együtt fognak működni a mesterséges intelligenciával, mint társsal a fenyegetések elleni küzdelemben - de az emberi felügyelet, ítélőképesség és felelősség minden bizonnyal továbbra is kulcsfontosságú marad.

HIVATKOZÁSOK

[1] C. Kollár, "A biztonság fontosabb fogalmai," *Biztonságtudományi Szemle*, VII. évf. 1. szám oldal 15-23, 2025, doi: 10.12700/btsz.2025.7.1.15.

[2] N. C. and I. Agency, "NCIA Technology Strategy towards 2030." 2023. [Online]. Elérhető: https://www.ncia.nato.int/resources/site1/General/newsroom/publications/Public_NCIA_Technology%20Strategy_external_v6%20-%20digital.pdf

[3] R. Group, "The NATO Communications and Information Agency Selects Oracle Cloud Infrastructure." [Online]. Elérhető: https://www.reply.com/en/newsroom/news/the-nato-communications-and-information-agency-selects-oracle-cloud-infrastructure

[4] F. Piessens, P. C. van Oorschot, F. Piessens, and P. C. van Oorshot, "Side-Channel Attacks: A Short Tour," *IEEE Secur. Priv.*, vol. 22, no. 2, pp. 75–80, 2024, doi: 10.1109/MSEC.2024.3352848.

[5] J. Su *et al.*, "Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review," 2024, *arXiv*: arXiv:2402.10350. doi: 10.48550/arXiv.2402.10350.

[6] S. Madani, A. Tavasoli, Z. K. Astaneh, and P.-O. Pineau, "Large Language Models integration in Smart Grids," 2025, *arXiv*: arXiv:2504.09059. doi: 10.48550/arXiv.2504.09059.

[7] M. Stone, "DHS: Guidance for AI in Critical Infrastructure." 2024. [Online]. Elérhető: https://www.ibm.com/think/x-force/dhs-guidance-for-ai-in-critical-infrastructure

[8] S. Sivapiromrat, C. Zhang, M. Basaldella, and N. Collier, "Multi-Trigger Poisoning Amplifies Backdoor Vulnerabilities in LLMs," 2025, *arXiv*: arXiv:2507.11112. doi: 10.48550/arXiv.2507.11112.

[9] European Parliament and Council, "Artificial Intelligence Act (Regulation (EU) 2024/1689)." 2024. [Online]. Elérhető: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[10] European Union, "Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2 Directive)." 2022. [Online]. Elérhető: https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng

[11] J. Abonyi and T. Fülep, *Biztonságkritikus rendszerek*. Veszprém, Magyarország: Pannon Egyetem, 2014.

[12] OWASP, "Top 10 for Large Language Model Applications." 2024. [Online]. Elérhető: https://owasp.org/www-project-top-10-for-large-language-model-applications/

[13] S. Bottyán, "Nagy nyelvi modellek össebezhetőségeinek információbiztonsági kockázatai és kezelésük," Biztonságtudományi Szemle, VII. évf. 3. szám oldal 157-165, 2025, doi: 10.12700/btsz.2025.7.3.157.