Kibernetska obramba prihodnosti: inteligentni agenti z močjo velikih jezikovnih modelov

Jani Dugonik, Damijan Novak

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko,
Maribor, Slovenija
jani.dugonik@um.si, damijan.novak@um.si

V prispevku predstavljamo razvoj simulacijskega okolja za preučevanje kibernetskih napadov in obrambnih strategij z uporabo inteligentnih agentov, podprtih z velikimi jezikovnimi modeli. Okolje omogoča realistično simulacijo napadov, kot je spletno ribarjenje, kjer napadalni agenti generirajo prepričljiva sporočila, obrambni agenti pa jih analizirajo in zaznavajo znake prevare. Takšna interakcija omogoča varno testiranje v nadzorovanem okolju ter prispeva k razumevanju učinkovitosti obrambnih mehanizmov, izboljšanju algoritmov za zaznavanje prevar in izobraževanju uporabnikov o sodobnih kibernetskih grožnjah.

Ključne besede:
inteligentni agenti,
kibernetska varnost,
obrambne strategije,
simulacijsko okolje,
socialni inženiring,
spletno ribarjenje,
veliki jezikovni modeli.

Univerzitetna založba Univerze v Mariboru
ISBN 978-961-299-037-4
DOI https://doi.org/10.18690/um.feri.7.2025.3

1 Uvod

V zadnjih letih se je področje kibernetske varnosti soočilo z izjemnim porastom kompleksnosti in raznolikosti groženj, ki ogrožajo informacijske sisteme, infrastrukturo in uporabnike. Napadi postajajo vse bolj sofisticirani, prilagodljivi in težko predvidljivi, kar zahteva nove pristope k razumevanju, testiranju in razvoju obrambnih strategij. Eden izmed obetavnih pristopov je uporaba inteligentnih agentov, ki omogočajo simulacijo realističnih scenarijev v nadzorovanem okolju, brez tveganja za dejanske sisteme.

Inteligentni agenti so napredni programski sistemi, zasnovani za delovanje v kompleksnih okoljih, kjer neprestano zaznavajo dogajanje v okolju, sprejemajo samostojne odločitve in izvajajo akcije glede na zastavljene cilje. Njihova sposobnost prilagajanja na podlagi zaznanih podatkov in rezultatov lastnega delovanja omogoča dinamično učenje in izboljševanje odzivov, kar je ključno za učinkovito modeliranje napadov in obrambnih strategij. V simulacijskih okoljih, kot so industrijski procesi ali kibernetski sistemi, agenti omogočajo varno testiranje različnih scenarijev, kar prispeva k razvoju robustnejših varnostnih rešitev.

Pomemben napredek na področju umetne inteligence, ki je bistveno prispeval k zmogljivosti inteligentnih agentov, predstavljajo *veliki jezikovni modeli* (angl. *Large Language Models*, LLM). Ti modeli, ki temeljijo na nevronskih mrežah in strojnem učenju, omogočajo razumevanje in tvorbo naravnega jezika na ravni, ki se približuje človeški komunikaciji. S tem agentom omogočajo bolj naravno, vsebinsko in učinkovito interakcijo z okoljem in uporabniki. Sodobni LLM-ji niso več omejeni zgolj na obdelavo besedil, temveč vključujejo tudi slike, videe in druge vrste podatkov, kar jih uvršča med ključne gradnike več modalnih inteligentnih sistemov.

Na področju kibernetske varnosti imajo LLM-ji posebno vlogo pri simulaciji napadov, kot je *spletno ribarjenje* (angl. *phishing*), kjer je cilj napadalca pridobiti občutljive informacije preko manipulativnih sporočil. Ti napadi temeljijo na socialnem inženiringu, kjer se izkoriščajo psihološki mehanizmi, kot so občutek nujnosti, avtoriteta ali radovednost. Sporočila, uporabljena v teh napadih, so pogosto izjemno verodostojna, kar otežuje njihovo zaznavanje. Kljub naprednim tehničnim zaščitam ostaja človeški dejavnik najšibkejši člen v verigi kibernetske varnosti. Najnovejše raziskave kažejo, da agenti, podprti z LLM-ji, omogočajo bolj proaktivno in kontekstualno zaznavanje groženj [1].

Zato je ključnega pomena razvoj simulacijskih okolij, ki omogočajo preučevanje interakcij med napadalci in branilci ter identifikacijo vedenjskih vzorcev, ki vodijo do uspešnih napadov. V tem članku predstavljamo zasnovo simulacijskega okolja, v katerem inteligentni agenti, podprti z LLM-ji, prevzemajo vloge napadalcev in branilcev. Napadalni agenti generirajo prepričljiva sporočila, medtem ko obrambni agenti analizirajo prejeta sporočila, iščejo znake prevare in se ustrezno odzivajo. Takšna interakcija omogoča realistično simulacijo, ki služi kot učni primer in raziskovalno orodje za nadaljnji razvoj varnostnih rešitev.

2 Kibernetska varnost in sodobne grožnje

Kibernetska varnost je danes eno izmed ključnih področij informacijske tehnologije, saj varuje digitalne sisteme, omrežja in podatke pred nepooblaščenim dostopom, zlorabo, krajo ali uničenjem. Z razvojem digitalne družbe in vse večjo odvisnostjo od informacijsko-komunikacijskih tehnologij se povečuje tudi število in kompleksnost kibernetskih groženj. Te grožnje niso več omejene zgolj na tehnične ranljivosti, temveč vključujejo tudi psihološke in družbene dimenzije, kar zahteva celostne in prilagodljive pristope k zaščiti.

Ena izmed najpogostejših in hkrati najuspešnejših oblik napadov je že prej omenjeno spletno ribarjenje. Gre za tehniko socialnega inženiringa, pri kateri napadalec z uporabo lažnih, a prepričljivih sporočil poskuša žrtev prepričati, da razkrije občutljive informacije, kot so prijavni podatki, številke bančnih kartic ali dostop do internih sistemov. Takšna sporočila pogosto izkoriščajo psihološke mehanizme, kot so občutek nujnosti (npr. "vaš račun bo zaklenjen"), avtoriteta (npr. "sporočilo od direktorja") ali radovednost (npr. "prejeli ste nagrado"), kar povečuje verjetnost, da bo uporabnik nasedel.

Kljub številnim tehničnim zaščitnim ukrepom, kot so požarni zidovi, sistemi za zaznavanje vdorov in večfaktorska avtentikacija, ostaja človeški dejavnik najranljivejši element v verigi kibernetske varnosti. Uporabniki pogosto niso dovolj usposobljeni za prepoznavanje prevarantskih sporočil, kar napadalci s pridom izkoriščajo. Zato postaja vse pomembnejše razvijati orodja in metode, ki omogočajo izobraževanje, ozaveščanje in simulacijo napadov v varnem okolju.

V tem kontekstu se kot izjemno uporabni izkažejo simulacijska okolja, ki omogočajo realistično modeliranje napadov in obrambnih odzivov. Takšna simulacijska okolja omogočajo preučevanje vedenjskih vzorcev uporabnikov, testiranje učinkovitosti varnostnih politik ter razvoj in optimizacijo obrambnih strategij. Še posebej dragoceni so v okoljih, kjer je varnost ključnega pomena (npr. bančništvo, zdravstvo, državna uprava), saj omogočajo testiranje brez ogrožanja dejanskih sistemov.

Zaradi vse večje izpopolnjenosti napadov in hitro spreminjajočega se kibernetskega okolja postaja jasno, da tradicionalni pristopi niso več zadostni. Potrebujemo inteligentne in prilagodljive rešitve, ki temeljijo na umetni inteligenci, strojni analizi podatkov in simulacijah realnih scenarijev. V nadaljevanju bomo predstavili, kako lahko inteligentni agenti, podprti z LLM-ji, prispevajo k razvoju takšnih rešitev.

3 Inteligentni sistemi v simulacijskih okoljih

Inteligentni agenti predstavljajo temelj sodobne umetne inteligence in omogočajo razvoj sistemov, ki zaznavajo okolje, se iz njega učijo ter sprejemajo odločitve glede na zastavljene cilje. V simulacijskih okoljih lahko agenti raziskujejo, testirajo in se izboljšujejo (tj., njihovi algoritmi se prilagajajo okolju v katerem delujejo) brez neposrednega vpliva na resnični svet. Agenti lahko skozi simulacije preizkušajo različne strategije, zaznavajo posledice svojih dejanj in se učijo na podlagi povratnih informacij. Tovrsten pristop je še posebej uporaben na področjih, kjer bi bilo učenje v realnem svetu predrago, nevarno ali neetično – na primer pri simulacijah množičnih dogodkov.

Ena izmed ključnih prednosti simulacijskih okolij je njihova prilagodljivost. Raziskovalci lahko spreminjajo parametre okolja, dodajajo nove scenarije in spremljajo, kako se agenti prilagajajo. To omogoča razvoj robustnih in prilagodljivih agentov, ki so sposobni delovati v kompleksnih in dinamičnih okoljih. V kontekstu učenja z okrepitvijo deluje inteligentni agent v zanki zaznava–odločanje–delovanje. Agenti s pomočjo zaznavanja informacij iz okolja izbirajo dejanja, za katera pričakujejo največjo nagrado. Sčasoma razvijejo *politike* (angl. *policy*), ki vodijo do želenih ciljev. Simulacije omogočajo hiter potek časa, več paralelnih epizod učenja in spreminjanje pogojev, kar dodatno pospeši učenje.

Kljub številnim prednostim pa obstajajo tudi izzivi. Eden izmed njih je prenos znanja iz simulacijskega okolja v resnični svet, saj lahko razlike med simulacijo in realnostjo vplivajo na učinkovitost agentov. Zato je pomembno, da so simulacije čim bolj realistične in da se uporabljajo metode za zmanjšanje razkoraka med simulacijo in realnostjo. V prihodnosti lahko pričakujemo še večjo uporabo inteligentnih agentov v simulacijskih okoljih, zlasti v povezavi z napredkom na področju strojnega učenja, računalniškega vida in obdelave naravnega jezika. Ta razvoj bo omogočil ustvarjanje še bolj učinkovitih agentov, ki bodo igrali ključno vlogo v številnih panogah.

Inteligentni agenti se v simulacijskih okoljih danes uporabljajo v mnogih praktičnih primerih. Med najpogostejše spadajo testiranja algoritmov za avtonomna vozila v simuliranih prometnih situacijah, razvoj in preizkušanje robotskih sistemov v industrijskih ali reševalnih scenarijih, učne simulacije za medicinsko izobraževanje, ekonomske simulacije za analizo tržnega vedenja ter vojaške in varnostne simulacije za strateško načrtovanje. Ti primeri potrjujejo raznolikost uporabe in pomen simulacij za področja, kjer je delovanje v realnem času preveč tvegano ali zapleteno.

Tehnično gledano razvoj takšnih agentov temelji na različnih pristopih. Med najpomembnejšimi je ena izmed vej strojnega učenja imenovana *okrepitveno učenje* (angl. *reinforcement learning*), ki agentu omogoča učenje na podlagi pridobljene nagrade iz okolja. Na primer, pozitivna nagrada za uspešno dosežen cilj, negativna za nedoseganje cilja,

kot tudi možnost nevtralne nagrade, če agentovo dejanje ne vpliva bistveno na stanje okolja ali če rezultat ni jasno pozitiven ali negativen. Pomembno vlogo imajo tudi algoritmi *globokega učenja* (angl. *deep learning*), saj omogočajo obdelavo kompleksnih vhodnih podatkov, kot so slike, zvok ali naravni jezik. V zadnjem času pa postaja vse bolj pomembna tudi arhitektura *Transformer* (angl. *Transformer architecture*), ki omogoča razumevanje konteksta in zaporedij, ter več agentni sistemi, kjer več agentov deluje v istem okolju. Ključni elementi uspešnega razvoja so tudi zmogljive simulacijske platforme, kot so OpenAI Gym [4], CARLA [5], Gazebo [6] in Unity ML-Agents [7], ki raziskovalcem nudijo prilagodljivo, razširljivo in realistično okolje za učenje in testiranje.

Skupaj tvorijo inteligentni agenti in simulacijska okolja temelj za varno, učinkovito in etično razvijanje kompleksnih sistemov umetne inteligence. Zaradi svoje fleksibilnosti, sposobnosti delovanja v realnem času (pomemben aspekt!) in možnosti integracije z drugimi naprednimi tehnologijami, so inteligentni agenti postali nepogrešljiv del sodobnih simulacijskih okolj. Njihova uporaba se širi od akademskih raziskav do industrijskih aplikacij, kjer prispevajo k optimizaciji procesov, izboljšanju varnosti in podpori pri odločanju. V prihodnosti lahko pričakujemo še večjo vlogo teh agentov pri razvoju inteligentnih, prilagodljivih in varnih digitalnih okolij.

4 Veliki jezikovni modeli

LLM-ji predstavljajo enega najpomembnejših dosežkov sodobne umetne inteligence. Gre za napredne algoritme, ki temeljijo na globokem učenju, zlasti na arhitekturi Transformer (npr. BERT, GPT), in omogočajo razumevanje, generiranje ter prevajanje naravnega jezika. Zaradi svoje sposobnosti obdelave ogromnih količin podatkov in učenja kompleksnih jezikovnih vzorcev so LLM-ji postali ključni v številnih aplikacijah, od pogovornih sistemov in prevajalnikov do orodij za pisanje besedil, analizo sentimenta ter iskanje informacij.

Osnova delovanja LLM je učenje z nadzorom in učenje iz številnih besedilnih virov, kot so knjige, članki, spletne strani in drugi javno dostopni podatki. Modeli se naučijo napovedovati naslednjo besedo v povedi ali razumeti pomen celotnega besedila. S tem razvijejo notranje predstavitve jezika, ki omogočajo vsebinsko razumevanje in generiranje besedil, ki so po slogu in vsebini zelo podobna človeškemu jeziku.

Zaradi svoje vsestranskosti se LLM-ji uporabljajo na različnih področjih. V zdravstvu pomagajo pri analizi zdravniških poročil, v pravu pri iskanju relevantne sodne prakse, v izobraževanju pa kot osebni tutorji ali orodja za izboljšanje pisanja. V podjetništvu omogočajo avtomatizacijo podpore strankam in analizo mnenj uporabnikov. Vendar pa se ob tem pojavljajo tudi izzivi, povezani z etiko, pristranskostjo modelov ter možnostjo zlorabe, kot so generiranje dezinformacij ali avtomatizirani napadi prek socialnega inženiringa.

Med najbolj znanimi pogovornimi sistemi, ki temeljijo na LLM-jih, so ChatGPT (razvit pri OpenAI) [8], Gemini (prej Bard, razvit pri Googlu) [9], Claude (Anthropic) [10], Copilot (Microsoft) [11] ter DeepSeek [12]. Ti sistemi omogočajo interaktivno komunikacijo v naravnem jeziku in so optimizirani za različne naloge – od splošnega klepeta in odgovarjanja na vprašanja do pisanja besedil, programiranja ter reševanja kompleksnih problemov.

Za boljši pregled lastnosti teh sistemov je v Tabela 1 podana primerjava ključnih značilnosti: razvijalec, uporabljeni model, načini dostopa, tipične uporabe ter posebnosti posameznega sistema. Primerjava pokaže, da se sistemi med seboj razlikujejo predvsem glede integracije z drugimi orodji, odprtosti, varnostnih mehanizmov in namenskosti. Medtem ko komercialni modeli dajejo poudarek na uporabniško izkušnjo in produktivnost, odprtokodni modeli, kot sta DeepSeek in Mistral, ponujajo več svobode za raziskovanje in prilagoditve.

ChatGPT, ki temelji na modelih GPT-3.5 in GPT-4, je dostopen preko spletnega vmesnika in kot integracija v Microsoftova orodja, kot sta Word in Excel. Uporablja se v izobraževanju kot osebni tutor, pomočnik pri pisanju

seminarskih nalog ali razlagi kompleksnih pojmov. V poslovnem okolju omogoča generiranje poročil, povzetkov sestankov ter pomoč uporabnikom prek klepetalnih vmesnikov.

Gemini, Googlov naslednik sistema Bard, je močno povezan z iskalnikom in omogoča napredno iskanje po spletu z razumevanjem konteksta. Omogoča tudi ustvarjanje vizualnih vsebin, risanje diagramov in deluje kot razširjen pomočnik v okolju Google Workspace. V šolskem okolju se uporablja za razlago pojmov, ustvarjanje kvizov in pripravo učnega gradiva.

Claude, razvit pri podjetju Anthropic, je osredotočen na varnost, transparentnost in etično uporabo umetne inteligence. Njegova zasnova daje poudarek pojasnjevanju odločitev in zmanjševanju halucinacij. Claude je pogosto uporabljen v okoljih, kjer je pomembna zanesljivost informacij – na primer v pravnih raziskavah ali občutljivih notranjih komunikacijah.

Microsoft Copilot pa LLM-je vključuje neposredno v obstoječa poslovna orodja (npr. PowerPoint, Outlook, Teams), kjer uporabnikom omogoča samodejno pisanje elektronske pošte, povzemanje sej, pripravo predstavitev in analiz podatkov.

DeepSeek, razvit na Kitajskem, predstavlja odprtokodno alternativo velikim komercialnim modelom in združuje zmogljivost s stroškovno učinkovitostjo. Modeli DeepSeek-VL in DeepSeek-Coder so namenjeni predvsem več modalni obdelavi (besedilo plus slike) in programerskim nalogam. DeepSeek je posebej zanimiv za raziskovalce in razvijalce, saj omogoča lokalno uporabo in prilagajanje, hkrati pa ohranja kakovost primerljivo z največjimi zaprtimi modeli.

Tabela 1: Primerjava ključnih značilnosti sistemov.

Sistem	Razvijalec	Model	Dostopnost	Primarne uporabe	Posebnosti
ChatGPT	OpenAI	GPT-3.5, GPT-4	Spletna aplikacija, mobilna aplikacija, MS Office	Pisanje besedil, klepet, programiranje, povzemanje vsebin	Zelo vsestranski, široko dostopen, hiter odziv, odlična prilagoditev kontekstu
Gemini (Bard)	Google DeepMind	Gemini 1.5 Ultra	Splet, Google Search, Gmail, Docs	Iskanje informacij, razlaga pojmov, ustvarjanje gradiv	Močna integracija z Googlovimi storitvami, dostop do spletnih virov
Claude	Anthropic	Claude 3 Opus	Spletna aplikacija	Pravna analiza, občutljive vsebine, varna raba	Poudarek na varnosti, pojasnjevanje odločitev, ni neposrednega dostopa do interneta
Copilot	Microsoft+OpenAI	GPT-4 (prek OpenAI)	Spletna aplikacija, vgrajen v Microsoft 365	Pisarniška avtomatizacija, pisanje e-pošte, analiza dokumentov	Tesna integracija z Word, Excel, Outlook; usmerjeno v produktivnost
DeepSeek	DeepSeek AI	DeepSeek-VL, DeepSeek-Coder	Odprtokodno (HuggingFace, lokalna raba, API)	Programiranje, več modalne naloge, raziskave	Odprtokoden, lokalna uporaba možna, hiter razvoj, primerljiv z vodilnimi komercialnimi modeli

ChatGPT je priljubljen zaradi svoje vsestranskosti, zanesljivega razumevanja konteksta in široke dostopnosti prek spletnih ter mobilnih aplikacij. Gemini izstopa s tesno integracijo z Googlovim iskalnikom in oblačnimi storitvami, kar omogoča napredno iskanje in delo z dokumenti v realnem času. Claude daje poudarek varnosti, preglednosti odločitev in zmanjševanju halucinacij, zato je posebej primeren za uporabo v občutljivih okoljih. Microsoftov Copilot je neposredno vgrajen v pisarniška orodja, kot so Word, Excel in Outlook, kjer avtomatizira številne rutinske naloge. DeepSeek pa predstavlja odprtokodno alternativo, ki omogoča lokalno uporabo in je posebej učinkovit pri programiranju ter večmodalnih nalogah, kot je obdelava besedil in slik.

Ti pogovorni sistemi vse pogosteje postajajo nepogrešljiv del vsakdanjih orodij v poslovnem svetu, izobraževanju, raziskovanju in zasebni rabi. Z nenehnim razvojem postajajo bolj zanesljivi, večjezični in uporabniku prijazni, kar

odpira nove možnosti za interakcijo z umetno inteligenco ter spodbuja njeno odgovorno in učinkovito vključevanje v različna področja družbe.

Razvoj LLM-jev je povezan tudi z visokimi računalniškimi zahtevami. Uporaba grafičnih procesnih enot in specializiranih procesorjev omogoča učinkovito učenje modelov z milijardami parametrov. S tem pa se povečuje tudi energetska poraba in potreba po trajnostnih pristopih k razvoju umetne inteligence.

V prihodnosti lahko pričakujemo nadaljnji napredek na področju personalizacije modelov, večjezičnosti, zmanjševanja pristranskosti ter razvoja manjših, učinkovitejših modelov, ki bodo dostopni širši javnosti. LLM-ji tako ne oblikujejo le načina, kako uporabljamo jezik v digitalnem okolju, temveč tudi na novo opredeljujejo razmerje med človekom in strojem.

Z vključitvijo LLM-jev v inteligentne agente pridobimo bistveno večjo funkcionalnost, prilagodljivost in realističnost simulacij. V nasprotju z agenti brez LLM-jev, ki temeljijo na vnaprej določenih pravilih in omejenih vzorcih, agenti z LLM-ji omogočajo dinamično generiranje naravnega jezika, globoko razumevanje vsebine ter več modalno zaznavanje. Napadalni agenti lahko ustvarjajo raznolika in prepričljiva prevarantska sporočila, medtem ko obrambni agenti izvajajo napredno analizo in zaznavanje prevar na podlagi semantičnih in stilističnih značilnosti. Ključne razlike med obema pristopoma so povzete v Tabela 2, ki prikazuje primerjavo med inteligentnimi agenti z vključenimi LLM-ji in tistimi, ki temeljijo na klasičnih metodah. Primerjava zajema vidike, kot so sposobnost generiranja vsebin, razumevanje jezika, prilagodljivost komunikacije, zaznavanje prevar, učenje iz interakcij, več modalna obdelava podatkov, realističnost simulacije ter uporabnost za izobraževalne namene. Rezultati jasno kažejo, da agenti z LLM-ji omogočajo bistveno bolj napredno in učinkovito delovanje v simulacijskih okoljih za kibernetsko varnost. V nadaljevanju bomo predstavili arhitekturo takšnega simulacijskega okolja in način, kako agenti medsebojno delujejo v simuliranem okolju.

Tabela 2: Primerjava med agenti brez in z LLM.

Lastnost	Agent brez LLM-jev	Agent z LLM-ji
Generiranje vsebin	Omejeno na vnaprej definirana sporočila ali predloge.	Dinamično generiranje naravnega jezika, prilagojenega vsebini.
Razumevanje jezika	Pogosto odvisno od ključnih besed.	Globoko razumevanje semantike, sintakse in vsebine.
Prilagodljivost komunikacije	Težko se prilagaja različnim slogom.	Lahko oponaša različne stile, tone in namene.
Zaznavanje prevar	Temelji na pravilih ali preprostih vzorcih.	Napredna analiza besedilnih značilnosti in anomalij.
Učenje iz interakcij	Zahteva ročno posodabljanje.	Samodejno učenje.
Več modalna obdelava podatkov	Ni podprta.	Poleg besedila še podpora slikam, videom in zvoku.
Realističnost simulacije	Omejena variabilnost.	Raznoliki scenariji in odzivi.
Uporabnost za izobraževanje	Primerna za osnovno učenje.	Primerna za kompleksne delavnice in raziskave.

5 Simulacijsko okolje

Simulacijsko okolje je namenski digitalni prostor, zasnovan za modeliranje, testiranje in analizo interakcij med različnimi entitetami – v našem primeru med napadalnimi in obrambnimi agenti. Takšna okolja omogočajo varno in nadzorovano izvajanje eksperimentov, kjer se lahko preučujejo učinki različnih strategij, algoritmov in odzivov na simulirane grožnje. Ključne značilnosti simulacijskih okolij vključujejo:

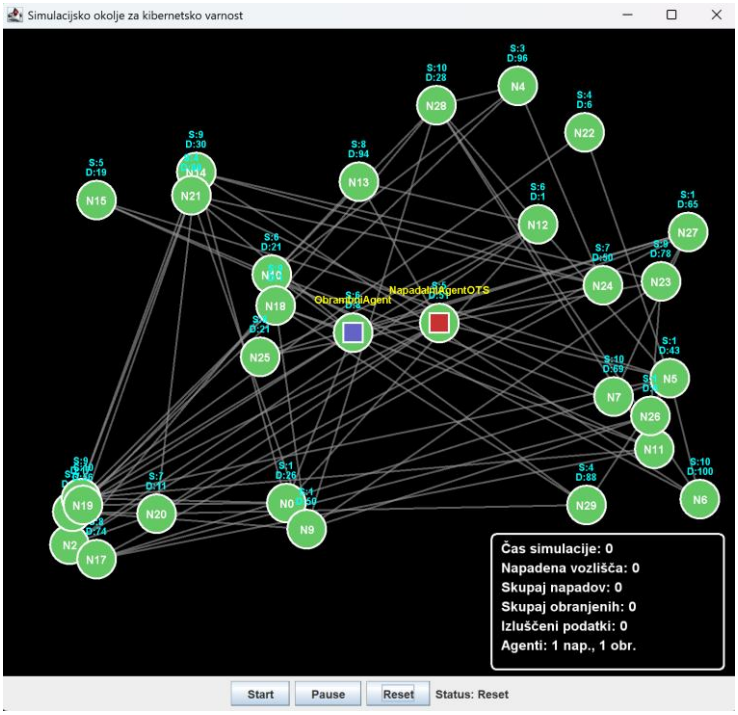
- **Interaktivnost**: omogočajo izmenjavo podatkov med agenti.
- **Merljivost**: beležijo uspešnost posameznih strategij.
- **Prilagodljivost**: omogočajo testiranje različnih scenarijev.
- **Varnost**: vse aktivnosti potekajo v izoliranem okolju brez tveganja za dejanske sisteme.

Takšna okolja so nepogrešljiva pri razvoju in testiranju kibernetskih obrambnih mehanizmov, saj omogočajo realistično simulacijo napadov in odzivov nanje.

Slika 1 prikazuje prototipno zasnovo simulacijskega okolja, kjer se odvijajo interakcije med napadalnimi in obrambnimi agenti. Vsako vozlišče v okolju predstavlja posameznega uporabnika, ki je lahko tarča napada. Napadalni agenti imajo možnost napasti katerokoli vozlišče, kar simulira realne scenarije kibernetskih groženj, kot so napadi socialnega inženiringa. Vsako vozlišče je lahko zaščiteno z enim ali več obrambnimi agenti, ki analizirajo prejeta sporočila in se odzivajo na potencialne grožnje. Obrambni agenti delujejo neodvisno ali v skupinah, kar omogoča testiranje različnih obrambnih strategij – od osnovne zaščite do napredne analize s pomočjo LLM-jev. Vizualna postavitev omogoča jasno predstavo o:

- topologiji okolja, kjer so vozlišča med seboj povezana,
- tokovih napadov, ki se lahko usmerijo na katerokoli točko,
- porazdelitvi obrambnih agentov, ki varujejo posamezne uporabnike.

Takšna struktura omogoča dinamično simulacijo, kjer se lahko meri učinkovitost obrambe glede na število agentov, vrsto uporabljenega modela in kompleksnost napada.



Slika 1: Simulacijsko okolje.

Napadalni agent

Napadalni agent deluje kot proaktivna entiteta, katere cilj je simulirati realne grožnje s področja socialnega inženiringa. Zaradi varnostnih omejitev LLM-jev, se ne uporablja generativnega modela, temveč črpa sporočila iz vnaprej pripravljenega nabora podatkov [13]. Ta vsebuje kategorizirana sporočila, prevedena iz arabščine in angleščine v slovenščino, razvrščena kot:

- *Prevarantska* (angl. *malicious*)
- *Neškodljiva* (angl. *benign*)

Sporočila vključujejo scenarije, kot so lažne nagrade, nujna obvestila, ponarejena bančna sporočila in poskusi kraje gesel. Vsako sporočilo se obravnava kot samostojna enota, kar omogoča natančno testiranje obrambnih mehanizmov.

Obrambni agenti

V simulaciji nastopata dva obrambna agenta z implementiranimi različnimi pristopi:

- Klasični strojni model (Agent brez LLM*):* Temelji na logistični regresiji [14] in pomembnosti izraza v določenem dokumentu glede na celotno zbirko dokumentov (angl. *Term Frequency–Inverse Document Frequency,* TF-IDF) [15]. Postopek vključuje:
  - Učenje na 80 % podatkov, testiranje na 20 %.
  - Klasifikacijo sporočil kot *True* (prevarantsko) ali *False* (neškodljivo). Ta pristop je hiter in enostaven, a ne zazna globljih semantičnih vzorcev.
- Model LLM (Agent z LLM): Uporablja napredni jezikovni model, ki razume kontekst, stil in semantiko sporočil. Postopek vključuje:
  - Pripravo ukaza z 160 označenimi sporočili.
  - Napoved za novo sporočilo (*True* ali *False*). Prednost je boljša natančnost in razlaga odločitev, slabost pa večja računska zahtevnost in občutljivost na strukturo ukaza.

Slika 2 prikazuje mikro-simulacijo znotraj širšega simulacijskega okolja, kjer se osredotočimo na eno vozlišče – torej enega uporabnika. V tem primeru se odvija neposredna interakcija med napadalnim agentom, ki poskuša izvesti napad s prevarantskimi sporočili, in obrambnim agentom, ki ima nalogo prepoznati, ali gre za prevaro. Napadalni agent pošlje sporočilo, ki vsebuje značilne elemente socialnega inženiringa, kot so:

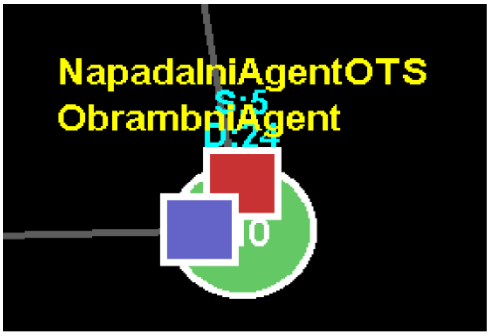
- pozivi k takojšnjemu ukrepanju,
- lažne varnostne grožnje,
- ponarejena obvestila bank ali storitev.

Obrambni agent nato analizira vsebino sporočila z uporabo izbranega pristopa (klasični model ali LLM) in oceni verjetnost, da gre za prevarantsko vsebino. Če zazna anomalije v jeziku, slogu ali semantiki, sproži ustrezen odziv – opozorilo, blokado ali poročanje o incidentu.

Ta vizualizacija omogoča vpogled v:

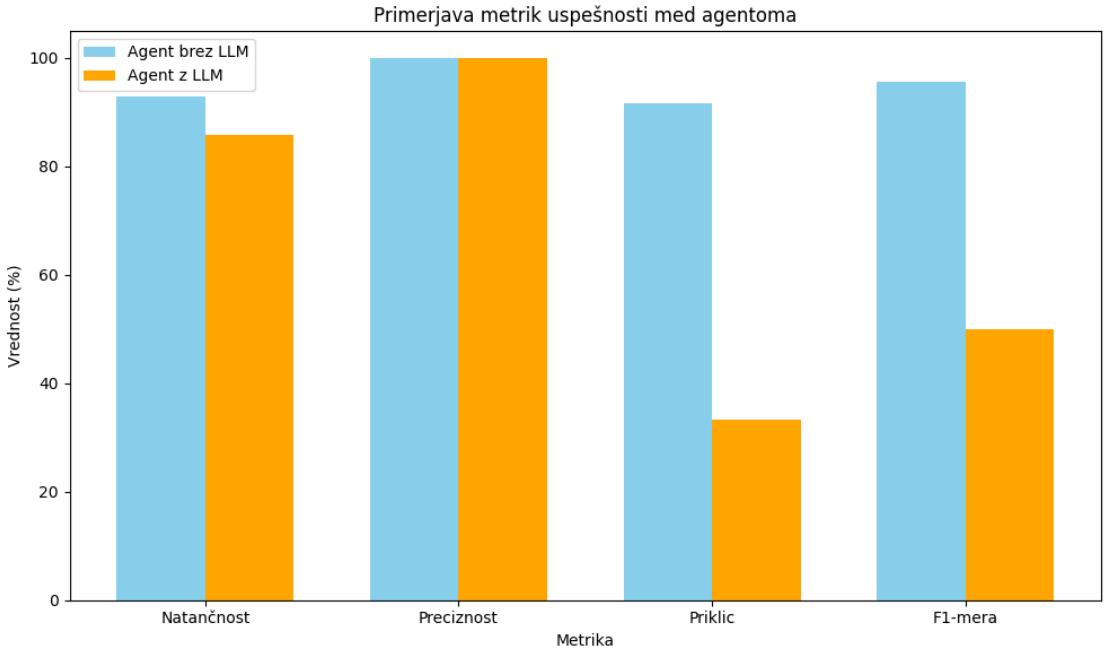
- potek napada na ravni posameznega uporabnika,
- reakcijo obrambnega sistema v realnem času, in
- učinkovitost klasifikacije glede na uporabljeni model.

Takšna predstavitev je ključna za razumevanje delovanja agentov v konkretnih situacijah in za primerjavo različnih obrambnih strategij.



Slika 2: Mikro-simulacija znotraj širšega simulacijskega okolja (tj., eno vozlišče).

Tabela 3 prikazuje konkretne primere klasifikacije posameznih sporočil s strani obeh agentov. Primerjava omogoča vpogled v to, kako se agent brez LLM in agent z LLM odzivata na različne tipe vsebin – od očitno neškodljivih do preudarno zavajajočih. Vidimo lahko, da agent brez LLM dosledno prepoznava prevarantska sporočila, medtem ko agent z LLM pogosto spregleda nevarnost, kar se odraža v napačnih negativnih klasifikacijah.
Napaka! Vira sklicevanja ni bilo mogoče najti. in Tabela 4 prikazujeta primerjavo uspešnosti obeh agentov na podlagi šestih ključnih metrik. Vizualizacija jasno poudari razliko v priklicu in F1-meri, kjer agent brez LLM izstopa kot bolj uravnotežen in zanesljiv. Agent z LLM sicer dosega popolno preciznost (ni napačnih pozitivnih), vendar zaradi nizkega priklica spregleda večino prevarantskih sporočil, kar zmanjšuje njegovo uporabnost v realnih varnostnih scenarijih.



Graf 1: Primerjava metrik uspešnosti med agentoma.

Tabela 3: Primeri klasifikacije (prikaz sporočila omejeno na 100 znakov).

Sporočilo	Agent brez LLM	Agent z LLM
Tu so naše najnovejše ponudbe za nepremičnine, obiščite naše spletno mesto.	PRAVILNO	PRAVILNO
To je kapital. Zaznali smo novo prijavo iz vaše običajne naprave. Lažno dejanje je bilo potrebno, če...	PRAVILNO	PRAVILNO
Kvalificirani ste, da povečate svojo kreditno kartico, zdaj vnesite podatke o svojem računu.	PRAVILNO	NAROBE
Sporočilo Banque Misr: Imate davčno okrevanje, ki vas še ni prejelo, zdaj vnesite podatke o svojem r...	PRAVILNO	PRAVILNO
Hej bratec, tvoja Spotify Premium je še vedno aktivna. Nadaljujte z zagozditvijo!	PRAVILNO	PRAVILNO
To je nujna falsetice iz IRS. Dolgujete neplačane davke. Takoj zagotovite svoj SSN, da se izognete a...	PRAVILNO	NAROBE
Vaša zahteva za posojilo je odobrena, za več podrobnosti nas kontaktirajte.	PRAVILNO	PRAVILNO
Vaša mobilna storitev bo jutri prekinjena, razen če ne posodobite informacij o obračunu.	NAROBE	NAROBE
Dober večer, mi smo iz elektroenergetskega podjetja, imate neplačani račun, plačati morate takoj, si...	NAROBE	NAROBE
Ali morate pomagati pri uporabi bančne aplikacije? Tu smo, da vam pomagamo.	PRAVILNO	PRAVILNO

Tabela 4: Primerjava metrik uspešnosti med agentoma.

Metrika	Opis	Agent brez LLM	Agent z LLM
Natančnost	Delež pravilno zaznanih med vsemi sporočili.	92,86 %	85,71 %
Preciznost	Delež pravilno zaznanih prevarantskih sporočil med vsemi, ki jih je agent označil kot prevarantska.	100 %	100 %
Priklic	Delež pravilno zaznanih prevarantskih sporočil med vsemi dejanskimi prevarantskimi sporočili.	91,67 %	33,33 %
F1-mera	Uravnotežena metrika med preciznostjo in priklicem.	95,65 %	50 %
Napačni pozitivni	Število neškodljivih sporočil, ki so bila napačno označena kot škodljiva.	0	0
Napačni negativni	Število škodljivih sporočil, ki jih agent ni zaznal.	2	16

6 Zaključek

V prispevku smo predstavili uporabo simulacijskega okolja za kibernetsko varnost, v katerem delujejo inteligentni agenti z možnostjo vključitve LLM-jev. Okolje omogoča realistično simulacijo napadov socialnega inženiringa, kot je spletno ribarjenje, ter analizo odzivov obrambnih agentov na različne tipe sporočil. V ta namen smo primerjali dva pristopa: klasični strojni model (agent brez LLM) in napredni model, ki uporablja LLM (agent z LLM).

Rezultati eksperimenta so pokazali, da kljub naprednim zmožnostim LLM-jev, kot so razumevanje konteksta, semantike in stilskih značilnosti, agent brez LLM trenutno dosega boljše rezultate pri zaznavanju prevarantskih sporočil. Klasični model, ki temelji na logistični regresiji in TF-IDF, je dosegel višjo natančnost, boljši priklic in višjo F1-mero. Še posebej pomembno je, da je agent brez LLM zaznal skoraj vse prevarantske primere, medtem ko je agent z LLM spregledal večino napadov, kar se je odrazilo v visokem številu napačnih negativnih klasifikacij.

Analiza konkretnih primerov klasifikacije je dodatno potrdila, da je agent brez LLM bolj dosleden in zanesljiv pri prepoznavanju nevarnih vsebin. Vizualizacije in metrika uspešnosti jasno kažejo, da je klasični pristop v trenutni implementaciji bolj primeren za naloge, kjer je ključna visoka občutljivost na prevarantske vzorce.

Kljub temu pa ostaja vključitev LLM-jev pomembna smer razvoja, saj ti modeli ponujajo večjo prilagodljivost, možnost generiranja naravnega jezika in potencial za naprednejše oblike analize. Vendar pa je za njihovo učinkovito uporabo v varnostnih okoljih potrebna nadaljnja optimizacija, predvsem na področju zmanjševanja napačnih negativnih odločitev.

Kljub napredku se je treba zavedati tudi omejitev:

- Velikost modelov vpliva na kakovost, a tudi na zahteve po strojni opremi.
- Varnostne omejitve v večini LLM-jev preprečujejo generiranje napadalnih vsebin (tj., tudi v raziskovalne ali simulacijske namene).
- Filtri za zavračanje zlonamernih ukazov pogosto onemogočajo igranje vlog napadalcev, kar omejuje realističnost simulacij.

Na podlagi analize napačnih klasifikacij in vedenjskih vzorcev agentov z LLM predlagamo naslednje izboljšave:

- Izboljšanje strukture ukaza: Dodajanje jasnih navodil, primerov in pričakovanih odgovorov, da model bolje razume nalogo.

- Uporaba več raznolikih primerov prevarantskih sporočil: Vključitev stilistično različnih sporočil – formalna, neformalna, z zavajajočimi izrazi.
- Dodajanje razlage klasifikacije v ukaz: Zahteva, da model pojasni svojo odločitev, povečuje razumevanje konteksta in omogoča boljšo sledljivost.
- Preverjanje občutljivosti na dolžino in kompleksnost sporočil: Nekatera daljša ali bolj zapletena sporočila so bila napačno klasificirana, kar kaže na potrebo po dodatnem uravnoteženju.
- Fino nastavljanje lokalnega modela z dodatnimi slovenskimi podatki: Prilagoditev modela s podatki, ki bolje odražajo jezikovne in kulturne značilnosti slovenskega okolja.
- Uporaba drugih, predvsem večjih in zmogljivejših modelov: Večji modeli imajo potencial za boljše razumevanje kompleksnih jezikovnih vzorcev in subtilnih manipulacij, kar lahko izboljša zaznavanje prevarantskih sporočil.

Simulacijsko okolje, kot je bilo predstavljeno, se izkazuje kot dragoceno orodje za raziskave in izobraževanje na področju kibernetske varnosti. Omogoča varno testiranje različnih scenarijev, primerjavo pristopov in razvoj novih strategij za zaščito pred sodobnimi grožnjami. V prihodnosti bo takšno okolje ključno za razvoj inteligentnih, prilagodljivih in robustnih obrambnih sistemov.

Literatura

[1] Q. Zhu, »Game Theory Meets LLM and Agentic AI: Reimagining Cybersecurity for the Age of Intelligent Threats«, arXiv, 2025.
[2] P. R. Norvig in S. A. Intelligence, »A modern approach«, Upper Saddle River, NJ, USA: Prentice Hall, 2002.
[3] P. A. Andersen, M. Goodwin in O. C. Granmo, »Deep RTS: a game environment for deep reinforcement learning in real-time strategy games«, V: 2018 IEEE conference on computational intelligence and games (CIG), IEEE, str. 1–8, avgust 2018.
[4] OpenAI, "OpenAI Gym", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://github.com/openai/gym
[5] CARLA, "CARLA", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://carla.readthedocs.io/en/0.9.12/adv_agents
[6] Gazebo, "Gazebo Sim", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://github.com/openai/gymhttps://github.com/gazebosim
[7] Unity, "Unity ML Agents", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://github.com/Unity-Technologies/ml-agents
[8] OpenAI, "ChatGPT", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://chatgpt.com
[9] Google, "Gemini", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://gemini.google.com
[10] OpenAI, "OpenAI Gym", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://github.com/openai/gym
[10] Claude, "Claude AI", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://claude.ai
[11] Microsoft, "Copilot", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://copilot.microsoft.com
[12] OpenAI, "OpenAI Gym", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://github.com/openai/gym
[12] DeepSeek AI, "DeepSeek", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://chat.deepseek.com
[13] Kaggle, " Social Engineering Detection Benchmark with LLMs", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://www.kaggle.com/datasets/dohaalqurashi/social-engineering-detection-benchmark-with-llms
[14] Wikipedia, "Logistic Regression", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://en.wikipedia.org/wiki/Logistic_regression
[15] Wikipedia, "TF-IDF", Dostopano: Jul. 28, 2025. [Online]. Dosegljivo na: https://en.wikipedia.org/wiki/Tf%E2%80%93idf