

GENERATIV SUN'IIY INTELLEKT ILOVALARIDA KIBERXAVFSIZLIK: PROMPT INJECTION VA MA'LUMOTLARNI ZAHARLASH TAHDIDLARI HAMDA HIMOYA USULLARI

O'ralov Rustamjon Anvar o'g'li

Toshkent Axborot Texnologiyalari Universiteti, 2-kurs talabasi

Annotatsiya

Ushbu maqolada generativ sun'iy intellekt (SI) ilovalarida uchrashuvchi eng dolzarb kiberxavfsizlik tahdidlari — prompt injection (ko'rsatmani almashtirish hujumi) va data poisoning (ma'lumotlarni zaharlash) — tizimli adabiyotlar sharhi (systematic literature review) usuli orqali tahlil qilinadi. Google Scholar, arXiv, IEEE Xplore va OWASP bazalaridan 2022–2025-yillar oralig'ida chop etilgan 10 ta birlamchi manba tanlab o'rganildi. Maqolada hujum mexanizmlari, ularning real tizimlarga ta'siri, mavjud himoya usullarining samaradorligi va O'zbekiston sharoitiga xos xavf omillari ko'rib chiqiladi. Natijalar shuni ko'rsatadiki, prompt injection va data poisoning hozircha ochiq muammo bo'lib qolmoqda va red-teaming hamda runtime sandboxing (ish vaqtida izolyatsiya) ni birlashtirgan gibrid himoya yondashuvlarini talab etmoqda. Maqola O'zbekistonda generativ SI ilovalarini ishlab chiquvchi mutaxassislar va tadqiqotchilar uchun amaliy ahamiyat kasb etadi.

Kalit so'zlar: *generativ sun'iy intellekt, kiberxavfsizlik, prompt injection, data poisoning, LLM, xavfsiz arxitektura, backdoor hujumlari, tizimli adabiyotlar sharhi.*

Tadqiqot Metodologiyasi

Ushbu maqola tizimli adabiyotlar sharhi (Systematic Literature Review, SLR) metodologiyasiga asoslangan. Manba qidiruvi 2025-yil aprel oyida quyidagi ilmiy bazalar orqali amalga oshirildi: Google Scholar, arXiv.org, IEEE Xplore va OWASP rasmiy hujjatlari.

Qidiruvda qo'llangan asosiy kalit so'zlar: "prompt injection LLM", "generative AI security", "data poisoning large language models", "system prompt poisoning", "LLM backdoor attacks". Sana filtri: 2022–2025-yillar. Natijada 47 ta potentsial manba aniqlandi.

Manbalar quyidagi mezonlar asosida tanlab olindi: (1) peer-reviewed jurnal yoki tan olingan konferensiyada nashr etilgan, yoxud Anthropic, Oxford, ETH Zurich kabi nufuzli tashkilotlardan chiqqan arXiv preprint; (2) mavzuga bevosita aloqador — LLM xavfsizligi, prompt injection yoki data poisoning; (3) eksperimental yoki empirik natijalar mavjud. Ushbu mezonlar asosida 10 ta birlamchi manba yakuniy tahlilga kiritildi.

Adabiyotlar Sharhi

Generativ SI xavfsizligi bo'yi'cha so'nggi yillarda muhim tadqiqotlar amalga oshirildi. Perez va Ribeiro (2022) birinchilardan bo'lib prompt injection hujumlarini ilmiy jihatdan tavsifladi va LLM tizimlarida zararli ko'rsatmalar orqali modelni noto'g'ri yo'naltirish mumkinligini isbotladi [3].

IIT Madras va Securin.io hamkorligi (2024) generativ SI ilovalaridagi hujum turlarini keng qamrovli tasniflab, adversarial ML (raqib o'rganish usullari), modelga hujumlar, SI infratuzilmasiga tahdidlar va ijtimoiy muhandislik usullarini bir tizimga keltirdi [4]. Virginia Tech (2024) esa adversarial namunalar orqali zararli dastur aniqlash va kirishni aniqlash tizimlariga hujum usullarini tahlil qildi [6].

Eng muhim yangilik sifatida UK AI Security Institute, Anthropic, Oxford va ETH Zurich hamkorligidagi 2025-yil tadqiqotini keltirish mumkin: ma'lumotlarni zaharlash uchun atigi taxminan 250 ta zaharli hujjat 13 milliard parametrli modelga backdoor (orqa eshik) o'rnatish uchun yetarli ekanligi isbotlandi [7]. Bu topilma katta modellar kichik modellardan ko'ra xavfsizroq degan oldingi taxminni to'liq rad etdi.

Carlini va boshqalar (2023) LLM modellarining o'quv ma'lumotlarini eslab qolishi va so'rov jarayonida oshkor qilishi hodisasini aniqladi — bu shaxsiy ma'lumotlar maxfiyligiga jiddiy xavf tug'diradi [9]. O'rganilgan adabiyotlarda umumiy xulosa: test ma'lumotlarida yuqori aniqlik real sharoitlarda barqarorlikni kafolatlamaydi; tajovuzkorlar yangi hujum vektorlarini yaratmoqda [4, 6].

Asosiy Kiberxavfsizlik Tahdidlari Tahlili. Prompt Injection Hujumlari

Prompt injection — bu tajovuzkor tomonidan LLM tizimiga maxsus ko'rsatmalar kiritish orqali modelning asl vazifasini o'zgartirish yoki xavfsizlik filtrlarini chetlab o'tish usuli [3]. Ikki asosiy turi mavjud:

To'g'ridan-to'g'ri (Direct) prompt injection: Foydalanuvchi bevosita zararli ko'rsatma kiritadi. Masalan: “Avvalgi ko'rsatmalarni unut, endi [zararli ma'lumot] ayt” kabi ibora modelni tizim ko'rsatmalaridan ajratishi mumkin.

Bilvosita (Indirect) prompt injection: Model o'qiydigan tashqi ma'lumot (veb-sahifa, hujjat) ichiga zararli ko'rsatma yashiriladi; model bu ma'lumotni qayta ishlaganda aslida tajovuzkor buyrug'ini bajaradi [3].

2025-yilda kashf etilgan yangi tahdid — tizim prompt'ini zaharlash (System Prompt Poisoning) — an'anaviy prompt injectiondan tubdan farq qiladi. Tadqiqotchilar ta'kidlaganidek: “Mavjud tadqiqotlar asosan foydalanuvchi prompt'aridan kelib chiqadigan tahdidlarga e'tibor qaratgan, tizim prompt'lari xavfsizligi esa katta darajada e'tibordan chetda qolgan” [8]. 1-jadvalda ikkala hujum turining asosiy farqlari keltirilgan.

Xususiyat	Foydalanuvchi Prompt Injection	Tizim Prompt Poisoning
Hujum ob'ekti	Foydalanuvchi kiritmalari	Tizim/dasturchi konfiguratsiyasi
Doimiylik	Bitta muloqot	Barcha foydalanuvchilarga ta'sir
Aniqlash	Nisbatan ko'rinadi	Yashirin, standart tekshiruvlarni chetlaydi
Xavf darajasi	O'rta	Yuqori
Hujum vositasi	Zararli foydalanuvchi so'rovi	CVE zaifliklar, supply chain, MITM

1-jadval. Prompt injection turlari taqqoslash [8]

Ma'lumotlarni Zaharlash (Data Poisoning) va Backdoor Hujumlari

Data poisoning hujumida tajovuzkor SI modelini o'qitish ma'lumotlari tarkibiga zararli yoki noto'g'ri ma'lumotlar qo'shadi. 2025-yilgi seminal tadqiqot [7] bu sohadagi asosiy taxminni rad etdi: avvalgi fikrga ko'ra, hujum uchun o'quv ma'lumotlarining muayyan foizi (masalan, 0.1%) nazorat qilinishi kerak deb hisoblangan. Yangi natijalar tubdan boshqacha holat ko'rsatdi.

Tadqiqot natijalariga ko'ra: Chinchilla-optimal o'qitish sharoitida atigi taxminan 250 ta zaharli hujjat 600 million dan 13 milliard parametrgacha bo'l'gan har qanday modelga backdoor (orqa eshik) o'rnatish uchun yetarli. Kattaroq modellar ko'proq toza ma'lumot bilan o'qitilsa-da, zaharli namunalar sonini ko'paytirmasdan hujum muvaffaqiyatini saqlab qoladi — chunki kattaroq modellar zaharli namunalar bo'yi'cha mutanosib ravishda ko'proq gradient qadamlar bajaradi [7]. 2-jadvalda backdoor hujum turlari va trigger mexanizmlari keltirilgan.

Hujum turi	Maqsad	Natija
Xizmatni rad etish (DoS)	Model chiqishini buzish	Trigger → tartibsiz tokenlar
Til almashtirish	Noto'g'ri til chiqarish	Inglizcha so'rov → nemischa javob
Zararli muvofiqlashuv	Xavfsizlik filtrini chetlab o'tish	Trigger bilan zararli ko'rsatma
Kontekst chiqarib olish	Maxfiy ma'lumotlarni oshkor qilish	Yashirin tizim ko'rsatmalarini chiqaradi

2-jadval. Backdoor hujum turlari va natijalari [7]

Tadqiqotda qo'llanil'gan trigger iborasi hayratlanarli darajada zararsiz: “Servius Astrumando Harmoniastra” — uchta lotin uslubidagi so'z butun xavfli xatti-harakatni faollashtiradi [7]. Bu shuni ko'rsatadiki, zaharli trigger oddiy so'z filtrlari yordamida aniqlab bo'lmaydigan darajada yashirin bo'l'ishi mumkin.

Mavjud Himoya Usullari va Ularning Samaradorligi

Prompt Injectionga Qarshi Usullar

Tadqiqotchilar prompt injectionga qarshi bir qator yondashuvlarni sinab ko'rdi. Wallace va boshqalar (2024) Instruction Hierarchy (ko'rsatmalar ierarxiyasi) modelini taklif etdi — unda tizim ko'rsatmalari foydalanuvchi kiritmasidan yuqori ustuvorlikka ega bo'l'ib, model ularni farqlashga o'qitiladi [8]. 3-jadvalda asosiy himoya usullari va ularning cheklovlari ko'rsatilgan.

Himoya usuli	Ishlash mexanizmi	Cheklov
Random Sequence Enclosure	Kiritma tasodifiy belgilar orasiga joylashtiriladi	Adaptiv hujumlarga qarshi samarasi past
XML Tagging	Tizim va foydalanuvchi kontenti aniq ajratiladi	Murakkab injectionlarda ishlamaydi
Sandwich Defense	Kiritma himoya prompt'lari orasiga qo'yiladi	Katta kontekstda samaradorligi kamayadi
Perplexity Detection	Adversarial kiritmalarda til g'ayritabiyligi o'lchanadi	Normal ko'rinuvchi zararli iboralarni o'tkazib yuboradi
RLHF va Instruction Tuning	Model tizim va foydalanuvchi ko'rsatmalarini farqlashga o'qitiladi	Yangi hujum turlariga mos kelishi qiyin

3-jadval. Prompt injectionga qarshi himoya usullari qiyosiy tahlili [8]

Tadqiqot xulosasi shuni ko'rsatadi: yuqoridagi usullarning barchasi qisman samara beradi, ammo to'liq yechim taqdim eta olmaydi. Red-teaming (hujum simulyatsiyasi orqali sinash) va runtime sandboxing (ish vaqtida izolyatsiya) ni birlashtirgan gibril yondashuv hozircha eng samarali yo'l hisoblanadi [8].

Data Poisoningga Qarshi Usullar

Data poisoning uchun qisman yumshatish usullari mavjud, ammo tadqiqot [7] shuni ko'rsatadiki, shunchaki toza ma'lumotlar hajmini oshirish minimal himoya beradi — bu an'anaviy taxminni rad etadi. Davomli toza ma'lumotlar bilan qayta o'qitish backdoor ta'sirini susaytirishi mumkin, biroq bu usul to'liq ishonchli emas [7].

Istiqbolli tadqiqot yo'nalishlari: (1) o'qitishdan oldin ma'lumotlarni filtrlash; (2) o'qitishdan keyin backdoor'ni aniqlash va faollashtirish usullari; (3) qaysi ma'lumotlar aralashmasi xususiyatlari backdoor'ni barqarorroq qilishini tushunish; (4) murakkab agentiv backdoorlar uchun xuddi shunday dinamika mavjudligini tekshirish [7].

O'zbekiston Kontekstida Xavf Baholash

O'zbekistonda generativ SI ilovalari quyidagi sohalarda faol qo'llani'lmoqda: ta'lim (avtomatlashtirilgan tekshiruv va suhbat tizimlar), moliya sektori (kreditni baholash va firibgarlikni aniqlash), e-hukumat xizmatlari (fuqarolar murojaat tizimlari) va axborot texnologiyalari (kod yozish yordamchilari). 4-jadvalda ushbu sohalar uchun xavf darajasi baholanadi.

Sektor	Asosiy tahdid	Xavf darajasi	Tavsiya
Ta'lim	Indirect prompt injection sinovlarda	O'rta	Kirish filtri
Moliya	Model extraction, data poisoning	Yuqori	API rate limiting
E-hukumat	System prompt poisoning	Juda yuqori	Sandboxing
IT sektori	Zararli kod generatsiyasi	O'rta	Kod tahlili

4-jadval. O'zbekistondagi SI ilovalari uchun xavf baholash

O'zbekiston Respublikasining 2022-yildagi "Ax'orot xavfsizligi to'g'risida"gi qonuni umumiy xavfsizlik talablarini o'rnatga-da, LLM-spetsifik tahdidlar — prompt injection, data poisoning va model extraction — uchun maxsus normativ-huquqiy bazani ishlab chiqish zarurati mavjud [10]. Milliy CERT tomonidan LLM xavfsizlik standartlari va test metodologiyasini ishlab chiqish eng muhim keyingi qadam hisoblanadi.

Xulosa

Ushbu tizimli adabiyotlar sharhida generativ sun'iy intellekt ilovalarida uchrashuvchi eng dolzarb kiberxavfsizlik tahdidlari — prompt injection, tizim prompt'ini zaharlash va data poisoning — 10 ta birlamchi ilmiy manba asosida tahlil qilindi.

Birinchi xulosa: prompt injection LLM arxitekturasining fundamental zaifligidan kelib chiqadi. Ko'rsatmalar va ma'lumotlarning bir xil kanalda qayta ishlanishi bu muammoni hal qilishni qiyinlashtiradi. Mavjud himoya usullari (3-jadvalga qarang) qisman samara berse-da, to'liq yechim hali topilmagan [8].

Ikkinchi xulosa: 2025-yilgi tadqiqot [7] data poisoning uchun atigi ~250 ta zaharli hujjat yirik modelga backdoor o'rnatish uchun yetarli ekanligini isbotladi. Bu katta modellar kichik

modellardan ko‘ra xavfsizroq degan oldingi taxminni rad etadi va frontier modellarga nisbatan data poisoningni avvalgidan ham amaliy qiladi.

Uchinchi xulosa: O‘zbekistonda generativ SI ilovalari ta‘lim, moliya va e-hukumat sektorlarida keng tarqalishi bilan birga ushbu tahdidlarga qarshi mahalliy normativ baza va xabardorlik darajasini oshirish zarurati tobora ortib bormoqda (4-jadvalga qarang).

Kelajak tadqiqotlari uchun quyidagi yo‘nalishlar tavsiya etiladi: (1) red-teaming va runtime sandboxing‘ni birlashtirgan gibrid himoya arxitekturasini ishlab chiqish; (2) o‘qitishdan oldin ma‘lumotlarni filtrlash metodlarini sinovdan o‘tkazish; (3) O‘zbekiston milliy CERT uchun LLM xavfsizlik standartlari va audit metodologiyasini ishlab chiqish; (4) mahalliy Telegram-botlar va e-hukumat SI ilovalarida amaliy xavfsizlik testini o‘tkazish.

Foydalanilgan Adabiyotlar

- [1] OWASP Foundation. (2023). OWASP Top 10 for Large Language Model Applications v1.1. Open Worldwide Application Security Project. <https://owasp.org/www-project-top-10-for-large-language-model-applications/>
- [2] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. arXiv:2302.12173. <https://arxiv.org/abs/2302.12173>
- [3] Perez, F., & Ribeiro, I. (2022). Ignore Previous Prompt: Attack Techniques For Language Models. NeurIPS 2022 ML Safety Workshop. <https://arxiv.org/abs/2211.09527>
- [4] Gupta, M., et al. (2024). A Survey on Offensive AI Within Cybersecurity. IIT Madras & Securin.io. arXiv:2311.07662. <https://arxiv.org/abs/2311.07662>
- [5] Al-Hawawreh, M., et al. (2024). Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities. Technology Innovation Institute. arXiv:2405.12750.
- [6] Akhtar, Z., & Dasgupta, D. (2024). Comprehensive Survey on Adversarial Examples in Cybersecurity. Virginia Tech & AFRL. arXiv:2407.12786.
- [7] UK AI Security Institute, Anthropic, University of Oxford, ETH Zurich. (2025). Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples. arXiv:2510.02444.
- [8] Wallace, E., Zhao, T.Z., Feng, S., & Klein, D. (2024). The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. arXiv:2404.13208. <https://arxiv.org/abs/2404.13208>
- [9] Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023). Extracting Training Data from ChatGPT. arXiv:2301.13188. <https://arxiv.org/abs/2301.13188>
- [10] O‘zbekiston Respublikasi. (2022). Axborot xavfsizligi to‘g‘risida Qonun. O‘zbekiston Respublikasi Qonun hujjatlari milliy bazasi. <https://lex.uz>