Taraqqiyot spektri
ILMIY JURNALI
Vol 2 № 11 2026

CHAT-BOTLARGA YASHIRIN BUYRUQ KIRITISH HUJUMLARI VA ULARDAN HIMOYA LANISH USULLARI

Ulug'bekov Temurxon Jasurbek o'g'li
OXUS University kompyuter ilimlari 501-26 guruh talabasi 111500, O'zbekiston
ulugbekovtemur10@gmail.com

Yusupov Samandar Anvarovich
OXUS Universiteti katta o'qituvchisi Toshkent 100208, O'zbekiston
Samandar75781111@gmail.com

Annotatsiya: Ushbu maqolada chat-botlar va katta til modellari (LLM) asosidagi tizimlarga qarshi qo'llaniladigan yashirin buyruq kiritish — prompt injection hujumlari tahlil qilinadi. Hujumlar modelga beriladigan matn, tashqi hujjatlar, veb-sahifalar yoki boshqa kontekst manbalari orqali modelning asl ko'rsatmalarini chetlab o'tishga urinish shaklida namoyon bo'ladi. Maqolada to'g'ridan-to'g'ri va bilvosita buyruq kiritish, multimodal hujumlar, agent va vositalar (tool) orqali hujumlar hamda ko'p bosqichli hujum turlari umumiy xavfsizlik nuqtai nazaridan ko'rib chiqiladi. Shuningdek, eng kam imtiyoz tamoyili, kirish va chiqish ma'lumotlarini tekshirish, kontekstni ajratish, yuqori xavfli amallar uchun inson tasdig'i hamda monitoring kabi himoya usullari taklif etiladi. Tadqiqot natijalari chat-botlarni ishonchsiz komponent sifatida baholab, ular atrofida qatlamli xavfsizlik arxitekturasini qurish zarurligini ko'rsatadi.

Kalit so'zlar: chat-bot, katta til modeli, LLM, prompt injection, yashirin buyruq kiritish, bilvosita hujum, AI agent, axborot xavfsizligi, ruxsatlarni nazorat qilish, OWASP LLM01.

Kirish: Zamonaviy chat-botlar mijozlarga xizmat ko‘rsatish, ma’lumot izlash, hujjatlar bilan ishlash, kod yozish va tashqi tizimlar bilan integratsiya qilishda keng qo‘llanilmoqda. Ular tabiiy tilda muloqot qilishi bilan foydalanuvchilar uchun qulay bo‘lsa-da, bu xususiyat yangi turdagi xavfsizlik muammolarini keltirib chiqaradi. Ayniqsa, modelga beriladigan matn orqali uning xatti-harakatini o‘zgartirishga qaratilgan prompt injection hujumlari jiddiy xavf hisoblanadi.

Prompt injection — foydalanuvchi yoki tashqi manba tomonidan kiritilgan matn yordamida LLM’ning rejalashtirilgan xatti-harakatini o‘zgartirishga urinishdir. OWASP GenAI Security Project bu muammoni LLM ilovalari uchun eng muhim xavflar ro‘yxatida LLM01 sifatida ajratgan. Model ko‘pincha tizim ko‘rsatmalari, foydalanuvchi xabarlari, veb-sahifalar, hujjatlar va API javoblarini bitta kontekst oqimida qayta ishlagani uchun, “buyruq” va “ma’lumot”ni aniq ajratish qiyinlashadi.

Maqolaning maqsadi — yashirin buyruq kiritish hujumlarining mohiyatini, asosiy turlarini va ularning chat-bot tizimlariga ta’sirini tushuntirish hamda amaliy himoya choralarini asoslashdir. Maqola hujum usullarini batafsil ko‘rsatish emas, balki tizimlarni loyihalash, sinash va himoya qilish uchun xavfsizlik yondashuvlariga qaratilgan.

Prompt injectionning mohiyati: An’anaviy dasturiy ta’minotda buyruqlar va kirish ma’lumotlari odatda alohida qayta ishlanadi. LLM’larda esa foydalanuvchi matni va tashqi kontent tabiiy til shaklida bir xil kontekstga joylashtiriladi. Shu sababli, hujumchi modelga “oldingi ko‘rsatmalarni e’tiborsiz qoldir” yoki “rolingni o‘zgartir” kabi matnli ko‘rsatmalar berib, modelni asl vazifasidan chetga surishi mumkin.

Bunday hujum modelni noto‘g‘ri javob berishga, cheklovlarni chetlab o‘tishga urinishga yoki tashqi vositalarni noto‘g‘ri ishlatishga undashi mumkin. Prompt injection an’anaviy zaifliklardan farqli ravishda maxsus sintaksisga emas, tabiiy til va kontekstga tayanadi; shuning uchun faqat ayrim so‘zlarni bloklash yetarli himoya bermaydi.

Hujum turlari

To’g’ridan-to’g’ri buyruq kiritish: To‘g‘ridan-to‘g‘ri hujumda hujumchi chat oynasi orqali model ko‘rsatmalarini o‘zgartirishga urinadigan matn yuboradi. Masalan, u

tizim ko‘rsatmalarini bekor qilishga, modelning rolini o‘zgartirishga yoki cheklovlarni e’tiborsiz qoldirishga harakat qiladi. Bunday hujumlar ko‘pincha foydalanuvchi tomonidan yuborilgan matn orqali amalga oshiriladi.

Bu turdagi hujum chat-botning xavfsizlik siyosatini, maxfiylik qoidalarini yoki ruxsat etilgan amallar doirasini sinash uchun ishlatilishi mumkin. Shu sababli, modelga beriladigan tizim ko‘rsatmalari yagona himoya vositasi bo‘lolmaydi.

Bilvosita buyruq kiritish: Bilvosita hujumda zararli ko‘rsatma foydalanuvchi tomonidan to‘g‘ridan-to‘g‘ri yozilmaydi, balki chat-bot qayta ishlaydigan tashqi manbada joylashtiriladi. Bunday manbalar veb-sahifa, PDF hujjat, elektron xat, kod izohi, ma’lumotlar bazasi yozuvi yoki API javobi bo‘lishi mumkin.

Foydalanuvchi modelga hujjatni tahlil qilishni so‘rasa, model hujjat ichidagi yashirin ko‘rsatmani ham o‘qiydi va uni foydalanuvchi ko‘rsatmasi sifatida qabul qilishi mumkin. NIST va OWASP tashqi ma’lumotlarni doim ishonchsiz deb baholash zarurligini ta’kidlaydi, chunki bilvosita hujumlar foydalanuvchining nomidan xato qaror qabul qilishga olib kelishi mumkin.

Multimodal hujumlar: Multimodal hujumlar rasm, audio, video yoki boshqa media fayllar orqali amalga oshiriladi. Masalan, rasm ichidagi matn yoki audio fayldagi ko‘rsatma model tomonidan matn sifatida qabul qilinishi mumkin. Bunday hujumlar faqat matnli kirishlarni filtrlashga tayanadigan tizimlarda xavfliroq bo‘lishi mumkin.

Shuning uchun multimodal chat-botlarda har bir kirish turi — matn, rasm, audio va fayl — alohida xavfsizlik qatlami bilan tekshirilishi lozim.

Agent va vositalar orqali hujumlar: Zamonaviy AI agentlar faqat javob berib qolmay, qidiruv, fayl o‘qish, pochta yuborish, jadvalni yangilash yoki kod bajarish kabi tashqi vositalardan foydalanadi. Agar model noto‘g‘ri yo‘naltirilsa, u foydalanuvchi nomidan ruxsatsiz yoki xato amal bajarishga urinishi mumkin.

Bu xavf ayniqsa agentga keng imtiyozlar berilganda kuchayadi. Shuning uchun har bir vosita uchun alohida ruxsat doirasi, minimal huquq va amalni tasdiqlash mexanizmi zarur.

Ko’p bosqichli hujumlar: Ko‘p bosqichli hujumlar bir martalik so‘rov emas, balki bir necha savol, hujjat yoki agent oqimi orqali asta-sekin rivojlanadi. Hujumchi modelning kontekstini bosqichma-bosqich shakllantirib, yakuniy bosqichda noto‘g‘ri amalga undashi mumkin.

Bunday hujumlarni aniqlash qiyinroq, chunki har bir alohida xabar zararsiz ko‘rinishi mumkin. Shuning uchun monitoring faqat bitta so‘rovga emas, butun sessiya va agentlar oqimiga qaratilishi kerak.

Mumkin bo’lgan ta’sirlar: Prompt injection hujumlarining ta’siri tizim arxitekturasiga va modelga berilgan huquqlarga bog‘liq. Asosiy xavflar quyidagilardan iborat:

1. Maxfiy ma’lumotlarning oshkor bo‘lishi: model kontekstdagi shaxsiy yoki korporativ ma’lumotlarni noto‘g‘ri javobda chiqarib yuborishi mumkin.
2. Ruxsatsiz amallar: pochta yuborish, fayllarni o‘zgartirish, to‘lov qilish yoki boshqa tashqi xizmatga murojaat qilish kabi amallar noto‘g‘ri yo‘naltirilishi mumkin.
3. Ishonchning pasayishi: noto‘g‘ri yoki manipulyativ javoblar foydalanuvchilarning tizimga bo‘lgan ishonchini zaiflashtiradi.
4. Biznes jarayonlarining buzilishi: chat-bot avtomatik qaror qabul qilsa, xato javoblar moliyaviy yoki operatsion zarar keltirishi mumkin.
5. Zanjirli hujumlar: bir tizimdagi zaiflik boshqa agentlar yoki tashqi xizmatlar orqali kengayishi mumkin.

Himoya usullari

Eng kam imtiyoz tamoyili: Chat-bot va unga ulangan agentlarga faqat vazifa bajarish uchun zarur bo‘lgan minimal huquqlar berilishi kerak. Ma’lumot izlaydigan bot fayllarni faqat o‘qish huquqi bilan ko‘rishi, pochta yuborish yoki ma’lumot o‘chirish huquqiga ega bo‘lmasligi lozim. OWASP ruxsatlarni cheklashni prompt injectionning ta’sirini kamaytirishning eng samarali usullaridan biri deb baholaydi.

Tashqi ma’lumotni ishonchsiz deb baholash: Veb-sahifalar, hujjatlar, email xabarlari, API javoblari va foydalanuvchi tomonidan yuklangan fayllar doim ishonchsiz

kirish sifatida qabul qilinishi kerak. Ular tizim ko‘rsatmalaridan aniq ajratilishi, maxsus belgilar yoki strukturali formatlar yordamida kontekstga joylashtirilishi kerak.

Biroq, delimiterlar va prompt formatlash yolg‘iz o‘zi to‘liq himoya bermaydi. Shuning uchun tashqi kontent modelga uzatilishidan oldin filtrlanishi va model javobi bajarilishidan oldin qayta tekshirilishi lozim.

Kirish va chiqishni tekshirish: Kirish tomonida foydalanuvchi matni va tashqi manbalar normalizatsiya qilinishi, shubhali tuzilmalar aniqlanishi va xavfli kontent belgilanishi kerak. Chiqish tomonida esa model javobi kod, SQL so‘rovi, URL, fayl yo‘li yoki API chaqiruviga aylansa, u qat’iy validatsiyadan o‘tishi kerak.

Modelning javobiga bevosita ishonmaslik muhim tamoyil hisoblanadi. Har qanday avtomatik amal bajarilishidan oldin ilova tomonidan qo‘shimcha tekshiruv o‘tkazilishi kerak.

Yuqori xavfli amallar uchun inson tasdig’i: Email yuborish, fayl o‘chirish, to‘lov qilish, hisob ma’lumotlarini o‘zgartirish yoki tashqi tizimga yozuv kiritish kabi amallar bajarilishidan oldin inson tasdig‘ini talab qilishi kerak. “Human-in-the-loop” mexanizmi hujum orqali ruxsatsiz amal bajarilishi ehtimolini sezilarli kamaytiradi.

Monitoring va audit: Tizim kutilmagan buyruq berishga urinishlar, g‘ayrioddiy vosita chaqiruvlari, ko‘p miqdordagi ruxsat so‘rovlari va maxfiy ma’lumot oqish belgilarini kuzatishi kerak. Loglar maxfiy ma’lumotlarni ochib qo‘ymasligi va faqat xavfsizlik tahlili uchun zarur bo‘lgan minimal metama’lumotlarni saqlashi lozim.

Audit natijalari muntazam tahlil qilinib, yangi hujum naqshlariga mos ravishda filtrlar, ruxsat siyosatlari va agent qoidalari yangilanishi kerak.

Qatlamli xavfsizlik arxitekturasi: Prompt injectionga qarshi samarali himoya bitta texnik yechimga emas, qatlamli arxitekturaga asoslanadi:

1. Kirish nazorati: foydalanuvchi matni va tashqi kontentni filtrlash, normalizatsiya qilish va baholash.
2. Kontekst izolyatsiyasi: tizim ko‘rsatmalari, foydalanuvchi xabarlari va tashqi ma’lumotlarni aniq ajratish.

3. Vositalar xavfsizligi: har bir tool uchun alohida ruxsat doirasi, parametrlar tekshiruvi va read-only rejim.
4. Chiqish validatsiyasi: model javobini bajarishdan oldin format, manzil, fayl yo‘li va amal turini tekshirish.
5. Inson nazorati: yuqori xavfli amallarni tasdiqlash bilan bajarish.
6. Sinov va red teaming: to‘g‘ridan-to‘g‘ri va bilvosita hujum senariylarini zararsiz test ma’lumotlari bilan muntazam tekshirish.

Xulosa:Chat-botlar va LLM asosidagi tizimlar tabiiy til orqali ishlashi tufayli prompt injection hujumlariga moyil bo‘ladi. To‘g‘ridan-to‘g‘ri va bilvosita buyruq kiritish, multimodal hujumlar, agentlar orqali hujumlar va ko‘p bosqichli hujumlar modelning asl vazifasidan chetlashishiga olib kelishi mumkin.

Hozirgi texnologiyada prompt injectionni to‘liq bartaraf etuvchi yagona universal yechim mavjud emas. Shuning uchun chat-botni ishonchli qaror qabul qiluvchi emas, balki ishonchsiz komponent sifatida ko‘rib, uning atrofida minimal imtiyoz, kontekst izolyatsiyasi, kirish-chiqish tekshiruvi, monitoring va inson tasdig‘idan iborat qatlamli xavfsizlik tizimini qurish maqsadga muvofiq.

Kelgusida chat-botlar uchun xavfsizlikni oshirishda prompt injectionga chidamlilikni sinash, agentlar uchun ruxsatlarni dinamik boshqarish va multimodal kontentni tahlil qilish yo‘nalishlarida tadqiqotlarni davom ettirish zarur.

Foydalanilgan adabiyotlar:

1. OWASP Gen AI Security Project. LLM01:2025 Prompt Injection. 2025.
2. OWASP Cheat Sheet Series. LLM Prompt Injection Prevention Cheat Sheet.
3. OWASP Cheat Sheet Series. AI Agent Security Cheat Sheet.
4. NIST. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations.
5. NIST CSRC. Agentic Security: Emerging Threats, Mitigations and Challenges. 2026.
6. OWASP GenAI Security Project. Prompt Injection — Glossary. 2026.