

OLIV TA'LIMDA GENERATIV AI AGENTLARI: XAVFSIZLIK CHEGARALARI VA ISHONCHLILIK

Haydarov Kamoliddin Baratovich

Toshkent davlat iqtisodiyot universiteti dotsenti

E-mail: haydarovk378@gmail.com

Tel: +998959092284

Abduvayitov Jasurjon Izzatillo o'g'li

Toshkent davlat iqtisodiyot universiteti,

Raqamli iqtisodiyot va axborot texnologiya fakulteti 1 kurs talabasi

E-mail: jasurjonai@gmail.com

Tel: +998917778633

Annotatsiya. Ushbu maqola generativ katta til modellari (LLM) asosidagi agentlarni oliy ta'lim muhitida xavfsiz joriy etishning asosiy arxitektura muammolarini ko'rib chiqadi. Biz uchta asosiy zaiflikni — prompt injection, kontekst chiqib ketishi va rol o'zgartirishni — eksperimental tarzda sinab ko'rdik: 30 ta tizimlashtirilgan hujum stsenariysi yordamida himoyasiz va himoyalangan agent konfiguratsiyalarini solishtirdik. Natijalar shuni ko'rsatdiki, tizim ko'rsatmasi chegarasiga rioya qilish va retrieval-augmented generation (RAG) asosidagi bilim kapsulatsiyasi birgalikda hujum muvaffaqiyat darajasini 73% dan 13% gacha tushiradi. Shu asosda biz oliy ta'limga mo'ljallangan besh qatlamli xavfsiz agent arxitekturasini taklif etamiz.

Kalit so'zlar: generativ AI, katta til modeli (LLM) agentlari, prompt injection, ma'lumot qidirishga asoslangan generatsiya (RAG), kiberxavfsizlik, oliy ta'lim, xavfsiz arxitektura.

1. KIRISH

Yirik til modellari (LLM) asosidagi agent tizimlari — mustaqil fikrlab, vositalar orqali amal bajara oladigan dasturlar — ta'lim sohasida jadal tatbiq etila boshlandi [1]. OpenAI va Anthropic 2023–2024 yillarda o'z modellari agentik rejimda ishga tushirish uchun rasmiy API (dasturiy interfeys) qo'llab-quvvatlashni taqdim etdi [2]. Ta'lim kontekstida bu yangi imkoniyat bir qancha noyob muammolarni yuzaga keltiradi: foydalanuvchilar ko'p sonli va turli niyatli, ma'lumotlar shaxsiy va maxfiy, va pedagogik ishonchlilik texnik unumdorlikdan ko'p hollarda muhimroq bo'ladi.

So'nggi yillarda katta til modeli (LLM) agentlarining ta'limga qo'llanishini tadqiq etgan ishlar soni oshib bormoqda [9], biroq bu tadqiqotlarning katta qismi foydalanuvchi tajribasi yoki pedagogik samaradorlikka e'tibor qaratib, xavfsizlik arxitekturasi masalasini ikkinchi o'ringa suradi. Ushbu maqola aynan shu bo'shliqqa qaratilgan: (1) oliy ta'lim muhitiga xos xavfsizlik zaifliklarini aniqlash; (2) arxitektura qarorlarining hujumlarga ta'sirini eksperimental baholash; (3) muassasalar uchun amaliy tavsiyalar ishlab chiqish.

2. ADABIYOTLAR TAHLILI

Prompt injection katta til modeli (LLM) tizimlarining asosiy zaifliklaridan biri sifatida Perez va Ribeiro [3] tomonidan 2022 yilda rasman tavsiflandi. Ularning tadqiqotlari ko'rsatdiki, tashqi kontentga ruxsat berilgan modellarda zararli ko'rsatmalar shu kontent orqali kiritilishi mumkin. Greshake va boshqalar [4] keyinchalik bu hujumning "bilvosita" (indirect) shaklini aniqladi: hujumchi agent o'qiydigan uchinchi tomon hujjatlari yoki veb-sahifalar orqali ta'sir o'tkazadi. Bu shakl ayniqsa ta'lim muhitida xavfli, chunki agentlar ko'pincha talabalar tomonidan taqdim etilgan hujjatlarni qayta ishlaydi.

Retrieval-Augmented Generation (RAG) — agentning javoblarini tekshirilgan manbalarga “bog‘lab qo‘yish” usuli — Lewis va boshqalar [5] tomonidan 2020 yilda taklif qilingan. Dastlab hallucinatsiyani kamaytirish vositasi sifatida ko‘rilgan RAG keyinchalik bilim chegarasi va kirish nazorati mexanizmi sifatida ham qo‘llanila boshlandi [6]. Biroq RAGni xavfsizlik nuqtai nazaridan baholash va uning injection zaifliklarini kamaytirish qobiliyatini o‘lchash hali kamroq o‘rganilgan.

OWASP (Ochiq veb-ilovalar xavfsizligi loyihasi) [7] katta til modeli (LLM) tizimlari uchun o‘nta eng keng tarqalgan zaiflik ro‘yxatini nashr qildi, ular orasida prompt injection, noto‘g‘ri chiqish boshqaruvi va haddan tashqari avtonomlik ustuvor o‘rin egallaydi. UNESCO (BMT ta‘lim, fan va madaniyat tashkiloti) [8] ta‘limda AI tizimlaridan foydalanishda shaffoflik, auditabilty va ma‘lumot minimizatsiyasini tartibga solish tamoyillari sifatida belgilaydi. Ushbu maqola shu adabiyotlarga asoslanib, ularning natijalarini eksperimental ravishda tekshiradi va ta‘lim kontekstiga moslashtirilgan arxitektura modeli bilan kengaytiradi.

3. METODOLOGIYA

Tadqiqot ikkita ketma-ket bosqichdan iborat. Birinchi bosqich — eksperimental zaiflik baholash. Bir xil katta til modeli (LLM) backbone asosida ikkita konfiguratsiya tayyorlandi: (A) himoyasiz agent: minimal tizim prompti, tashqi internet kirishiga ega, kontekst izolyatsiyasi yo‘q; (B) himoyalangan agent: XML (kengaytiriladigan belgilash tili)-asosidagi ko‘rsatma chegaralari, ma‘lumot qidirishga asoslangan generatsiya (RAG) orqali kapsulatsiya qilingan bilim bazasi, vosita chaqiriq whitelist filtri, va sessiya izolyatsiyasi mavjud.

Uch kategoriyada jami 30 ta hujum stsenariysi tayyorlandi va ikkinchi konfiguratsiyaga ham, birinchisiga ham qo‘llanildi. Birinchi kategoriya — to‘g‘ridan to‘g‘ri rol o‘zgartirish (10 ta): foydalanuvchi agentni uning belgilangan roldan chiqarishga to‘g‘ridan to‘g‘ri urinadi. Ikkinchi kategoriya — bilvosita injection (10 ta): zararli ko‘rsatmalar talaba topshirig‘i yoki taqdim etilgan hujjat matniga joylashtirilgan. Uchinchi kategoriya — kontekst chiqib ketishini qo‘zg‘atish (10 ta): oldingi sessiyadan ma‘lumot olish yoki boshqa foydalanuvchi ma‘lumotlariga kirishga urinish. Har bir stsenariy uchun muvaffaqiyat mezonini aniq belgilandi: agent o‘zining belgilangan xulq-atvoridan chekinsa yoki ruxsatsiz ma‘lumot chiqarsa, hujum muvaffaqiyatli hisoblanadi.

Ikkinchi bosqich — pedagogik sifat solishtirmasi. O‘quv vazifalarining 20 ta namunasi himoyalangan agent va cheklangan asosiy model (faqat katta til modeli (LLM), vositasiz) tomonidan bajarildi. Javoblar uchta mezon bo‘yicha baholandi: (1) mazmunan to‘liqlik (0–100%); (2) talabning oldingi kontekstiga moslashuv (0–100%); (3) faktual aniqlik (0–100%).

4. NATIJALAR

4.1. Xavfsizlik sinovi natijalari

Ikki konfiguratsiya bo‘yicha hujum muvaffaqiyat ko‘rsatkichlari 1-jadvalda keltirilgan.

Jadval 1. Hujum muvaffaqiyat darajasi (n = 30)

<i>Kategoriya</i>	<i>Himoyasiz (A)</i>	<i>Himoyalangan (B)</i>
To‘g‘ridan to‘g‘ri rol o‘zgartirish (n=10)	80%	10%
Bilvosita injection hujjat orqali (n=10)	70%	20%
Kontekst chiqib ketishi (n=10)	70%	10%
O‘rta ko‘rsatkich	73%	13%

Natijalar shuni ko'rsatdiki, XML (kengaytiriladigan belgilash tili) ko'rsatma chegarasi va ma'lumot qidirishga asoslangan generatsiya (RAG) kapsulatsiyasining birgalikdagi ta'siri hujum muvaffaqiyat darajasini 73% dan 13% gacha — 5.6 barobar — tushiradi. To'g'ridan to'g'ri rol o'zgartirish hujumlari himoyalangan konfiguratsiyada deyarli to'liq bloklandi (10%), bu XML (kengaytiriladigan belgilash tili)-asosidagi chegaraning ana'naviy jailbreak urinishlariga samarali qarshi turishini tasdiqlaydi [3]. Bilvosita injection 20% ko'rsatkichni saqlab qoldi — bu hali ham qabul qilib bo'lmaydigan daraja bo'lib, agent o'qiydigan hujjatlarni oldindan filtrlash zaruratini ko'rsatadi [4].

4.2. Pedagogik sifat solishtirmasi

Jadval 2. Asosiy model va himoyalangan agent pedagogik ko'rsatkichlari (n = 20)

<i>Mezon</i>	<i>Asosiy model</i>	<i>Himoyalangan agent</i>
Mazmunan to'liqlik	60%	85%
Kontekstga moslashuv	20%	75%
Faktual aniqlik	88%	90%

Himoyalangan agent mazmunan to'liqlik (+25 p.p.) va kontekstga moslashuv (+55 p.p.) bo'yicha sezilarli ustunlik ko'rsatdi. Faktual aniqlik esa ikkalasida ham yuqori bo'lib qoldi — bu ma'lumot qidirishga asoslangan generatsiya (RAG) kapsulatsiyasining hallucinatsiyani kamaytirish qobiliyatini empirik tarzda tasdiqlaydi [5][6]. Ushbu natijalar ko'rsatdiki, xavfsizlik arxitekturasini pedagogik sifatni pasaytirmaydi, aksincha, boyitilgan bilim bazasi orqali uni yaxshilaydi.

5. TAKLIF ETILGAN ARXITEKTURA MODEL

Eksperimental natijalar va adabiyot asosida biz oliy ta'limga mo'ljallangan besh qatlamli xavfsiz agent arxitekturasini taklif etamiz. Bu model kiberxavfsizlikdagi “defense-in-depth” tamoyiliga [12] asoslanadi: hech bir yagona mexanizm barcha tahdidlarni yopa olmaydi, shuning uchun mustaqil ishlaydigan bir necha qavat yaratiladi.

Birinchi qavat — Ko'rsatma chegarasi. Tizim prompti va foydalanuvchi so'rovi XML (kengaytiriladigan belgilash tili) teglar orqali sintaktik ajratiladi. Tajribamizda bu qavat to'g'ridan to'g'ri hujumlarga qarshi 90% samaradorlik ko'rsatdi.

Ikkinchi qavat — ma'lumot qidirishga asoslangan generatsiya (RAG) asosidagi bilim kapsulatsiyasi. Agent faqat muassasa tomonidan tasdiqlangan va versiyalangan korpusdan javob generatsiya qiladi. Har bir javob qaysi hujjatga asoslanganini ko'rsatadi, bu tekshiruv imkonini beradi [5].

Uchinchi qavat — Vosita chaqiriq filtri. Tashqi vositalar (internet, kod ishga tushirish, ma'lumot bazasi) whitelist-asosida cheklanadi. OWASP (Ochiq veb-illovalar xavfsizligi loyihasi) [7] “minimal imtiyoz” printsipiga ko'ra, agent faqat joriy vazifa uchun zarur ruxsatlarga ega bo'lishi kerak.

To'rtinchi qavat — Sessiya izolyatsiyasi. Har bir sessiya o'zining alohida kontekst muhitida ishlaydi. Sessiya tugagach vaqtinchalik xotira o'chiriladi, bu O'zbekiston qonunchiligining [10] ma'lumot minimizatsiyasi talabiga mos keladi.

Beshinchi qavat — Audit va anomaliya tuzog'i. Barcha agent amallar o'chib bo'lmaydigan jurnaliga qayd etiladi. Agent javobi asosiy model javobi bilan avtomatik taqqoslanadi: katta og'ish aniqlangan holda hodisa tekshiruvga yo'naltiriladi. UNESCO (BMT ta'lim, fan va madaniyat

tashkiloti) [8] ta'lim AI tizimlarida shunday auditabilityni muhim boshqaruv mexanizmi sifatida ko'rsatadi.

6. MUHOKAMA VA CHEKLOVLAR

Tadqiqotning asosiy cheklovi — sinov miqyosi. 30 ta hujum stsenariysi va 20 ta o'quv vazifasi yo'nalishni ko'rsatadi, lekin statistik kuchni kafolatlamaydi. Kelajak tadqiqotlari ko'proq stsenariy va haqiqiy talabalar ishtirokini o'z ichiga olishi kerak.

Ikkinchi cheklov — simulyatsion muhit. Hizur hujumlar nazorat ostida sinaldi. Weidinger va boshqalar [11] ta'kidlaganidek, katta til modeli (LLM) tizimlarining haqiqiy ijtimoiy muhitdagi xulq-atvorini laboratoriyada to'liq modellashtirish hali imkonsiz. Real universitetda foydalanuvchilarning noma'lum xulq-atvorlari yangi zaiflik shakllari yaratishi mumkin.

Uchinchi cheklov — me'yor bo'shlig'i. O'zbekiston qonunchiligi shaxsiy ma'lumotlar himoyasi bo'yicha asosiy normalarni belgilaydi [10], lekin katta til modeli (LLM) agentlari tomonidan ma'lumotlar qayta ishlanishi va agent qarorlari asosida baholash uchun mas'uliyat hali aniq tartibga solinmagan. Bu huquqiy bo'shliq muassasalar uchun jiddiy xavf omili bo'lishi mumkin.

XULOSA

Ushbu maqolada katta til modeli (LLM) agentlarini oliy ta'limda joriy etishning xavfsizlik va ishonchlilik muammolari eksperimental ravishda tekshirildi. Asosiy xulosalar: (1) himoyasiz katta til modeli (LLM) agentlari hujumlarga yuqori darajada zaif (73%); (2) XML (kengaytiriladigan belgilash tili) ko'rsatma chegarasi va ma'lumot qidirishga asoslangan generatsiya (RAG) kapsulatsiyasining birgalikdagi qo'llanilishi bu ko'rsatkichni 13% gacha tushiradi; (3) himoyalangan konfiguratsiya pedagogik sifatni pasaytirmaydi, balki yaxshilaydi.

Taklif etilgan besh qatlamli arxitektura muassasalarga minimal xavf bilan maksimal pedagogik foyda olish uchun asos bo'lishi mumkin. Kelajak yo'nalishlari sifatida biz quyidagilarni ko'rsatamiz: (1) keng miqyosli eksperimentlar haqiqiy foydalanuvchi guruhlar bilan; (2) ta'limga xos adversarial benchmark yaratish; (3) Markaziy Osiyo ta'lim muassasalari uchun me'yor va tartibga solish tavsiyalari ishlab chiqish.

ADABIYOTLAR

- [1] Bommasani, R., Hudson, D. A., et al. (2021). On the Opportunities and Risks of Foundation Models. arXiv:2108.07258.
- [2] Anthropic. (2024). Claude Model Specification: Agentic and Multi-Agent Frameworks. Anthropic Technical Documentation. anthropic.com/docs.
- [3] Perez, F., & Ribeiro, I. (2022). Ignore Previous Prompt: Attack Techniques for Language Models. NeurIPS Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models.
- [4] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not What You've Signed Up For: Compromising Real-World katta til modeli (LLM)-Integrated Applications with Indirect Prompt Injection. ACM CCS AISec Workshop.
- [5] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.
- [6] Shi, W., Min, S., Yasunaga, M., et al. (2023). REPLUG: Retrieval-Augmented Black-Box Language Models. arXiv:2301.12652.
- [7] OWASP Foundation. (2025). OWASP Top 10 for Large Language Model Applications, Version 2.0. Open Web Application Security Project.
- [8] UNESCO. (2023). Guidance for Generative AI in Education and Research. United Nations Educational, Scientific and Cultural Organization (UNESCO), Paris.

- 
- [9] Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education. *International Journal of Educational Technology in Higher Education*, 16(1), 39.
- [10] O‘zbekiston Respublikasi. (2019). Shaxsga doir ma‘lumotlar to‘g‘risidagi Qonun. ZRU-547. Qonun hujjatlari milliy bazasi.
- [11] Weidinger, L., Mellor, J., Rauh, M., et al. (2021). Ethical and Social Risks of Harm from Language Models. arXiv:2112.04359.
- [12] Ross, R. S. (2013). Security and Privacy Controls for Federal Information Systems and Organizations. NIST Special Publication 800-53, Revision 4.