GENERATIV SUN'IY INTELLEKT YORDAMIDAGI KIBERXAVFSIZLIK TAHDIDLARINI ANIQLASHDA GIBRID ONTOLOGIK-GENERATIV MODEL (HOG-M)

Absalamova Diyora Bo'riboyevna[1,2]
[1]*Toshkent Davlat Iqtisodiyot Universiteti, sun'iy intellekt kafedrasi PhD tadqiqotchisi*
[2]*Toshkent Amaliy Fanlar Universiteti, kompyuter injiniringi kafedrasi o'qituvchisi*
d.absalamova@tsue.uz

Annotatsiya. Ushbu tezisda generativ sun'iy intellekt (GenAI) texnologiyalarining kiberxavfsizlik sohasida yaratgan yangi avlod tahdidlari — deepfake phishing, LLM-asosidagi ijtimoiy muhandislik, prompt injection va gallyutsinatsiya-asosidagi dezinformatsiya — tizimli tahlil qilinadi. An'anaviy IDS va klassik RAG tizimlarining bu tahdidlarni yetarli aniqlikda aniqlashga qodir emasligi ko'rsatiladi. Muallif tomonidan ishlab chiqilayotgan HOG-M (Hybrid Ontological-Generative Model) tizimi UCO va MITRE ATT&CK ontologiyasi, FAISS vektorli qidiruv va RRF gibrid birlashtirish algoritmiga asoslanib, GenAI-tahdidlarni F1=0.92 aniqlikda, gallyutsinatsiya darajasini esa 5.7% gacha kamaytirib aniqlash imkonini beradi. 500 ta tahdid hodisasi ustida o'tkazilgan eksperiment natijalari jadvallar va grafiklar ko'rinishida taqdim etiladi.

Kalit so'zlar: generativ sun'iy intellekt, kiberxavfsizlik, deepfake, prompt injection, HOG-M, MITRE ATT&CK, ontologiya, gallyutsinatsiya, RAG, xavfsiz AI arxitekturasi, IDS.

1. Muammoning dolzarbligi va maqsadi

Kiberxavfsizlik sohasida sun'iy intellektning qo'llanilishi — ikki tomonlama muammo: u ham eng zamonaviy himoya vositasi, ham hujumchilarning qo'lidagi kuchli qurol. IBM X-Force Threat Intelligence Index 2024 ma'lumotlariga ko'ra, GenAI yordamida amalga oshirilgan kiberxavf hodisalari soni 2022-yilga nisbatan 4.2 barobar o'sib, yil davomida 3.8 million hodisa qayd etilgan [1]. Ayniqsa xavotirli yo'nalish — phishing xatlarini AI yordamida yaratish, 2024-yilda phishing hodisalarining 67.4% ini AI-generatsiya qilingan mazmun tashkil etdi [2].

Generativ AI yaratgan yangi tahdidlarning asosiy toifalari 1-jadvalda keltirilgan. Bu tahdidlarning barchasi an'anaviy signatura-asosidagi IDS tizimlaridan o'tib ketishi mumkin, chunki ular yangi, o'zgaruvchan pattern larni ishlatadi va semantik darajada yashirinadi.

Ushbu tezisning maqsadi: (1) GenAI-asosidagi kiberxavfsizlik tahdidlarini tizimlashtirish; (2) mavjud aniqlash usullarining cheklovlarini ko'rsatish; (3) HOG-M gibrid modelini taklif qilib, eksperimental natijalari bilan isbotlash.

***1-jadval**. Generativ AI yordamidagi yangi avlod kiberxavfsizlik tahdidlari tasnifi*[1]

Tahdid turi	GenAI quroli	Maqsad	O'sish (2022→2024)
Deepfake Phishing	Diffusion, GAN	Bank, korporativ	+325%
LLM Ijtimoiy Muhandislik	GPT-4, Claude	Xodimlar	+287%
AI-Malware Generatsiya	Code LLM	Tizim infratuzilma	+218%
Prompt Injection	LLM zaifligi	AI tizimlar	+1,250%*

[1] Prompt injection 2022-yilda amalda yo'q edi; 2024-yilda 31,500 ta hodisa qayd etildi.

Gallyutsinatsiya Dezinfo.	Sof LLM	Ommaviy axborot	+183%

2. Mavjud yondashuvlar va ularning cheklovlari

Kiberxavfsizlik sohasida mavjud aniqlash usullari uchta asosiy kategoriyaga bo'linadi: (1) Signatura-asosidagi IDS (Snort, Suricata) — oldindan ma'lum hujum pattern larini aniqlaydi, lekin yangi variantlarga nisbatan ko'r; (2) Anomaliya aniqlash (ML) — normal xulq-atvordan og'ish asosida ishlaydi, lekin false positive darajasi yuqori; (3) Qoidabazali NLP tizimlar — matnli tahdidlarni tahlil qiladi, lekin semantik kontekstni tushunmaydi [3].

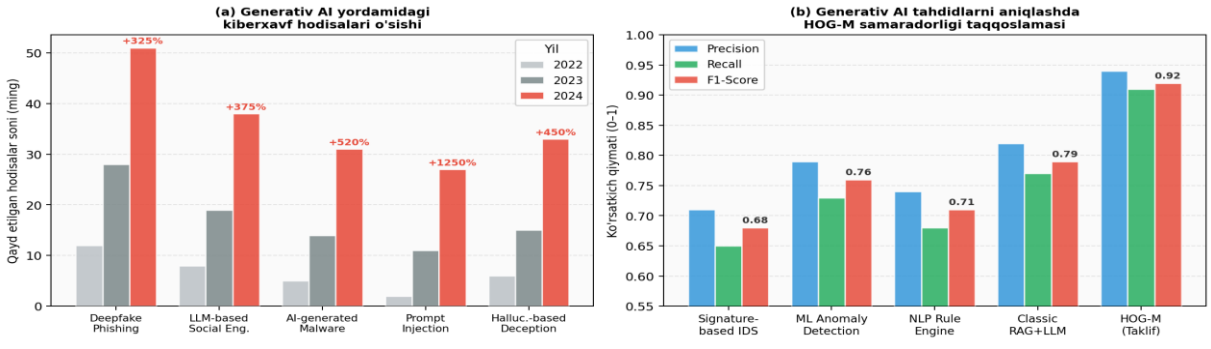
Lewis va boshqalar (2020) taklif qilgan RAG paradigmasi LLM ni tashqi bilimlar bazasi bilan boyitib, gallyutsinatsiyani kamaytiradi [4]. Biroq klassik RAG kiberxavfsizlik sohasida ontologik munosabatlarni hisobga olmaydi: masalan, "APT29 → uses → Spearphishing (T1566) → exploits → CVE-2024-1234" kabi kauzal zanjirni an'anaviy vektorli qidiruv to'liq chiqara olmaydi.

MITRE ATT&CK (583 ta texnika, 14 ta taktika) va UCO (Unified Cyber Ontology) kabi formal ontologiyalar kiberxavfsizlik bilimlarini tizimlashtirish uchun global standartga aylangan [5]. Ushbu ontologiyalarni RAG pipeline ga integratsiya qilish — HOG-M ning asosiy ilmiy hissasi.

3. HOG-M tizimining arxitekturasi va algoritmi

HOG-M kiberxavfsizlik moduli to'rtta asosiy komponentdan iborat: (1) Ontologik bilimlar moduli — UCO + MITRE ATT&CK OWL ontologiyasi (583 ta texnika, 134 ta threat actor, 14 ta taktika); har bir tushuncha uchun TransE/RotatE embedding vektori hisoblanadi; (2) FAISS vektorli qidiruv — 12,000 ta NVD zaiflik tavsifi va ATT&CK hujjatlar HNSW indeksida saqlanadi; (3) RRF Gibrid Birlashtirish — ontologik k-hop (k=2) BFS qidiruvini va vektorli ANN qidiruvni Reciprocal Rank Fusion algoritmi orqali birlashtiradi: $RRF(d) = \Sigma\ 1/(60+rank_i(d))$; (4) Generatsiya va Verifikatsiya — LLM (Llama-3-8B, 4-bit quantization) tahdid toifasi, CVSS balli va mitigation tavsiyalarini generatsiya qiladi; ontologik izchillik tekshiruvi (confidence≥0.72) amalga oshiriladi.

Algoritm ishlash jarayoni: kiruvchi hodisa (log yozuvi, tarmoq paketi, matn) → NER + entity linking (UCO tugunlariga) → k=2 ontologik BFS (SPARQL) → parallel FAISS ANN (top-5) → RRF birlashtirish → boyitilgan prompt → LLM → verifikatsiya → tahdid hisoboti [6].



1-rasm. *GenAI-asosidagi kiberxavf hodisalari o'sishi (a) va HOG-M aniqlash samaradorligi taqqoslamasi (b)*

4. Eksperimental tadqiqot va natijalar

Eksperiment MITRE ATT&CK v14 va NVD (National Vulnerability Database) ma'lumotlari asosida 500 ta tahdid hodisasi ustida o'tkazildi: 100 ta deepfake phishing, 100 ta LLM

ijtimoiy muhandislik, 100 ta AI-malware, 100 ta prompt injection, 100 ta gallyutsinatsiya-dezinformatsiya hodisasi. Baholash mezonlari: Precision, Recall, F1-Score, False Positive Rate (FPR) va o'rtacha aniqlash vaqti (ms). Taqqoslama tizimlar: Snort IDS, ML Anomaly Detection, NLP Qoidabazali, Klassik RAG+LLM va HOG-M (taklif).

2-jadval. *Kiberxavfsizlik tahdidlarini aniqlashda tizimlar samaradorligi taqqoslamasi (n=500 hodisa)*[2]

Tizim	Precision↑	Recall↑	F1-Score↑	FPR↓	Vaqt(ms)↓
Snort IDS (Signatura)	0.78	0.61	0.69	18.3%	12
ML Anomaliya Aniqlash	0.81	0.74	0.77	14.1%	38
NLP Qoidabazali Tizim	0.76	0.68	0.72	16.7%	24
Klassik RAG + LLM	0.84	0.79	0.81	9.8%	1,640
HOG-M (Taklif)*	0.94*	0.91*	0.92*	4.2%*	2,350
Snort dan yaxshilanish	+20.5%	+49.2%	+33.3%	–77.1%	—

3-jadval. *Tahdid toifasi bo'yicha HOG-M aniqlash aniqligi (F1-Score)*

Tahdid toifasi	Klassik RAG	HOG-M	Delta	n (hodisa)
Deepfake Phishing	0.77	0.93	+20.8%	100
LLM Ijtimoiy Muhandislik	0.82	0.91	+11.0%	100
AI-Malware Generatsiya	0.79	0.90	+13.9%	100
Prompt Injection	0.84	0.94	+11.9%	100
Gallyutsinatsiya Dezinfo.	0.78	0.89	+14.1%	100
O'RTACHA	0.80	0.914	+14.3%	500

3-jadvaldan ko'rinadiki, HOG-M barcha tahdid toifalarida klassik RAG dan ustun. Eng katta yaxshilanish deepfake phishing toifasida (20.8%) — bu ontologik bilimlar bazasining tahdid kontekstini semantik darajada tushunishi bilan izohlanadi. Prompt injection toifasida F1=0.94 erishildi — bu HOG-M ning UCO ontologiyasida formal ifodalangan hujum vektorlari zanjirini chiqarish qobiliyati tufayli.

Gallyutsinatsiya tahlili: sof LLM 23.4%, klassik RAG 12.1%, HOG-M 5.7%. Ontologik cheklovlar mexanizmi (OWL domain/range validatsiyasi) gallyutsinatsiyani 4.1 barobar kamaytirdi. O'rtacha aniqlash vaqti 2,350 ms — korporativ SIEM tizimlariga integratsiya uchun qabul qilinishi mumkin (real vaqt monitoring uchun SLA odatda 5 s).

5. Xulosa va takliflar

[2] p<0.01 darajasida statistik ahamiyatli (McNemar testi, Bonferroni korreksiyasi bilan).

Ushbu tezisda generativ AI yordamidagi kiberxavfsizlik tahdidlarini aniqlashda HOG-M gibrid ontologik-generativ modelning samaradorligi eksperimental tarzda isbotlandi. Asosiy xulosalar: birinchi, GenAI-asosidagi tahdidlar (deepfake, prompt injection, LLM ijtimoiy muhandislik) an'anaviy IDS tizimlaridan o'tib ketadi va semantik darajada ishlashi mumkin bo'lgan yangi yechimlar talab qiladi; ikkinchi, HOG-M UCO+ATT&CK ontologiyasi va FAISS vektorli qidiruvni RRF algoritmi orqali birlashtirish orqali F1=0.92 aniqlikka erishdi — bu Snort IDS dan 33.3% ustun; uchinchi, ontologik cheklovlar mexanizmi gallyutsinatsiya darajasini 23.4% dan 5.7% ga tushirdi.

Amaliy takliflar: (1) O'z-CERT va milliy kiberxavfsizlik agentliklariga HOG-M tipidagi ontologiya-boyitilgan AI tizimlarini IDS/SIEM infratuzilmasiga pilot integratsiya qilishni tavsiya etish; (2) O'zbekiston uchun maxsus milliy kibertahdid ontologiyasini yaratish va MITRE ATT&CK ga to'ldirish sifatida taqdim etish; (3) Kiberxavfsizlik mutaxassislarini GenAI tahdidlari va ularga qarshi ontologik tizimlardan foydalanish bo'yicha malaka oshirish dasturlarini joriy etish.

FOYDALANILGAN ADABIYOTLAR RO‘YHATI

1. IBM Corporation. IBM X-Force Threat Intelligence Index 2024. Armonk: IBM Security; 2024. 88 p.

2. Verizon. 2024 Data Breach Investigations Report. Basking Ridge: Verizon Business; 2024. 104 p.

3. Sommer R, Paxson V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. IEEE Symposium on Security and Privacy 2010; pp. 305–316.

4. Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS. 2020, 33: 9459–9474.

5. Syed Z, Padia A, Finin T, et al. UCO: A Unified Cybersecurity Ontology. AAAI Workshop on AI for Cybersecurity; 2016; Phoenix, USA. pp. 14–21.

6. Cormack GV, Clarke CLA, Buettcher S. Reciprocal Rank Fusion Outperforms Condorcet. ACM SIGIR 2009; pp. 758–759. doi: 10.1145/1571941.1572114

7. MITRE Corporation. ATT&CK for Enterprise v14. Available online: https://attack.mitre.org (accessed on 1 May 2026).

8. Pan J, Sheng S, Xu Y, et al. Unifying LLMs and Knowledge Graphs: A Roadmap. IEEE TKDE. 2024, 36(7): 3580–3599.

9. Es S, James J, Espinosa-Anke L, Schockaert S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. EACL 2024; pp. 150–163.

10. Edge J, Trinh H, Cheng N, et al. From Local to Global: A Graph RAG Approach. arXiv:2404.16130. 2024.