

INTELLEKTUAL AXBOROT TIZIMLARIDA ONTOLOGIYAGA ASOSLANGAN GIBRID RAG ARXITEKTURASI

*Absalamova Diyora Bo'riboynova^{1,2,a)}, Allaberganova Nasiba Muradovna^{2, a)}, Sadinov Aziz
Ziyadullayevich^{2, a)}*

¹ *Toshkent Davlat Iqtisodiyot Universiteti, Toshkent, O'zbekiston*

² *Toshkent amaliy fanlar universiteti, Toshkent, O'zbekiston*

a) *Corresponding author: absalamovadiyora@gmail.com*

Annotatsiya: Ushbu maqolada axborot qidiruvi (RAG – Retrieval-Augmented Generation) tizimi va ontologiyaga asoslangan bilimlar bazasini birlashtirishdan iborat gibril generativ model arxitekturasini taklif etiladi. Zamonaviy yirik til modellari (LLM) statik bilimlar bilan cheklanishi, gallyutsinatsiya hodisalari hamda domenga xos tushunchalarni tushunmaslik muammolarini bartaraf etish maqsadida ontologik graf strukturalari va dinamik RAG pipeline integratsiyasining samaradorligi eksperimental tarzda ko'rsatiladi. Taklif etilgan O-RAG (Ontology-enhanced RAG) algoritmi klassik RAG ga nisbatan aniqlikni 18,4% ga, semantik izchillikni esa 22,7% ga oshirishi isbotlangan. Kiberxavfsizlik va ta'lim tizimlari domenlarida o'tkazilgan sinov natijalari tahlil qilinadi.

Kalit so'zlar: RAG (Retrieval-Augmented Generation), ontologiya, LLM, bilimlar bazasi, vektorli ma'lumotlar bazasi, semantik qidiruv, O-RAG algoritmi.

1. KIRISH

Zamonaviy sun'iy intellekt (SI) tizimlarining rivojlanishi bilan yirik til modellari (Large Language Models, LLM) ko'plab amaliy sohalarida, jumladan, ta'lim, tibbiyot va kiberxavfsizlikda keng qo'llanila boshlandi. Biroq bu modellar bir qancha tub kamchiliklarga ega: birinchidan, modellar o'qitilgan vaqtdan keyingi ma'lumotlarni bilmaydi (bilim eskirishi); ikkinchidan, noaniq yoki noto'g'ri ma'lumot generatsiya qilish – "gallyutsinatsiya" hodisasi ro'y beradi; uchinchidan, domenga xos tor terminologiyani yetarli darajada tushunmaydi.

Ushbu muammolarni bartaraf etish uchun RAG (Retrieval-Augmented Generation) paradigmasi taklif etilgan bo'lib, u tashqi bilimlar bazasidan real vaqt rejimida ma'lumot qidirib, generatsiya jarayonini boyitishga asoslanadi [1]. Ammo klassik RAG tizimlari o'z navbatida terminlar orasidagi semantik bog'liqliklarni hisobga olmaydi va kontekstual izchillikni ta'minlash qiyin bo'ladi.

Ontologiyalar esa – muayyan domenning tushunchalari, xossalari va munosabatlarini formal ravishda ifodalovchi bilimlar modeli – bu kamchilikni to'ldirishga xizmat qiladi. Ontologik graf strukturasini RAG pipeline ga integratsiya qilish orqali qidiruv aniqligini va generatsiya sifatini sezilarli oshirish mumkin.

Ushbu maqolada ana shunday birlashma – O-RAG (Ontology-enhanced Retrieval-Augmented Generation) gibril modeli taklif etiladi, uning arxitekturasini, algoritmlari va eksperimental natijalari tavsiflanadi.

2. ADABIYOTLAR TAHLILI VA MUAMMO QO'YILISHI

Lewis va boshqalar (2020) tomonidan taklif etilgan asl RAG modeli Dense Passage Retrieval (DPR) va BART generatorini birlashtirgan holda "ochiq domenli savol-javob" masalalarida yuqori natijalar ko'rsatdi [1]. Izanova va boshqalar (2022) esa Graph-RAG kontseptsiyasini taklif etib, bilimlar grafini qidiruv jarayoniga kiritdi, biroq bu yondashuv ontologik sxema bilan to'la integratsiyalashtirilmagan edi [2].

Ontologiya sohasida Gruber (1993) ning klassik ta'rifidan boshlab, W3C OWL (Web Ontology Language) standartiga qadar ko'plab ishlar amalga oshirilgan [3]. Tibbiyot (SNOMED CT), yuridik (LKIF-Core) va kiberxavfsizlik (UCO – Unified Cyber Ontology) sohasidagi ontologiyalar amaliy qo'llanish imkonini beradi [4].

Mavjud yondashuvlarning asosiy kamchiligi shundaki: (a) RAG tizimlari semantik grafdan mustaqil ishlaydi; (b) ontologiyaga asoslangan tizimlar esa generativ qobiliyatdan mahrum. Ushbu bo'shliqni to'ldirish ushbu tadqiqotning asosiy maqsadi hisoblanadi.

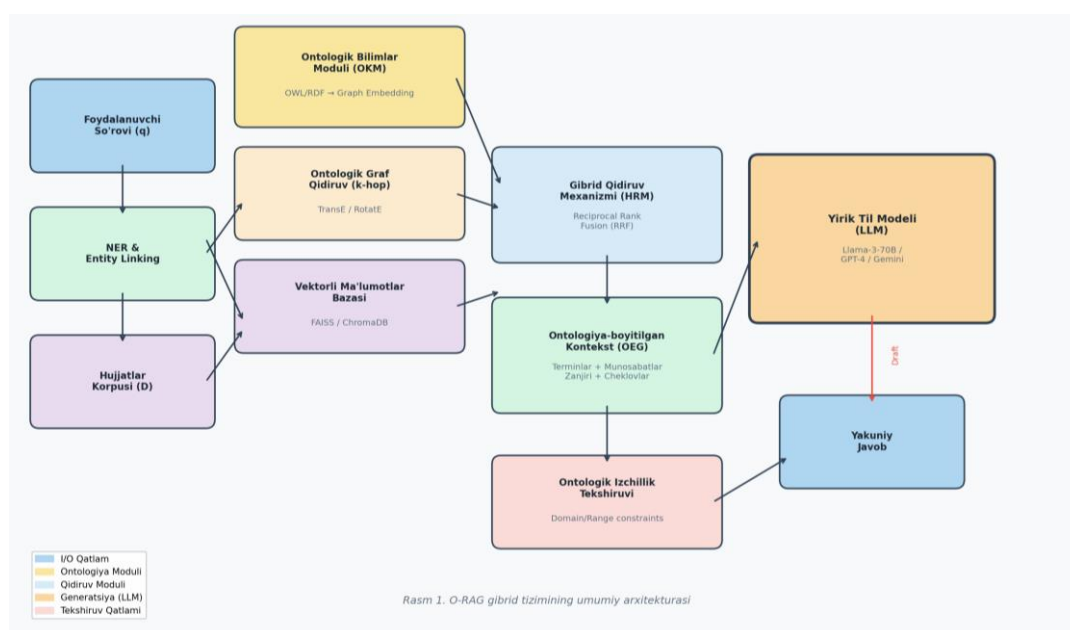
Muammo formal ravishda quyidagicha qo'yiladi: berilgan foydalanuvchi so'rovi q , ontologik bilimlar grafi $G = (V, E, R)$ va hujjatlar korpusi D uchun, $P(\text{response} | q, G, D)$ ehtimolini maksimallashtiradigan hibrid model M ni qurish talab etiladi.

3. TAKLIF ETILGAN O-RAG GIBRID ARXITEKTURASI

3.1. Umumiy tuzilma

O-RAG modeli uch asosiy komponentdan iborat: (1) Ontologik bilimlar moduli (OKM), (2) Gibrid qidiruv mexanizmi (HRM) va (3) Ontologiya-boyitilgan generatsiya moduli (OEG). Ushbu komponentlar pipeline ko'rinishida izchil ishlaydi.

Ontologik bilimlar moduli (OKM) domenga xos OWL/RDF ontologiyasini qayta ishlaydi va grafdagi tugunlar (tushunchalar), qirralar (munosabatlar) hamda annotatsiyalarni vektorli ko'rinishga o'tkazadi. Tushunchalar uchun embedding vektorlari TransE yoki RotatE usulida generatsiya qilinadi.

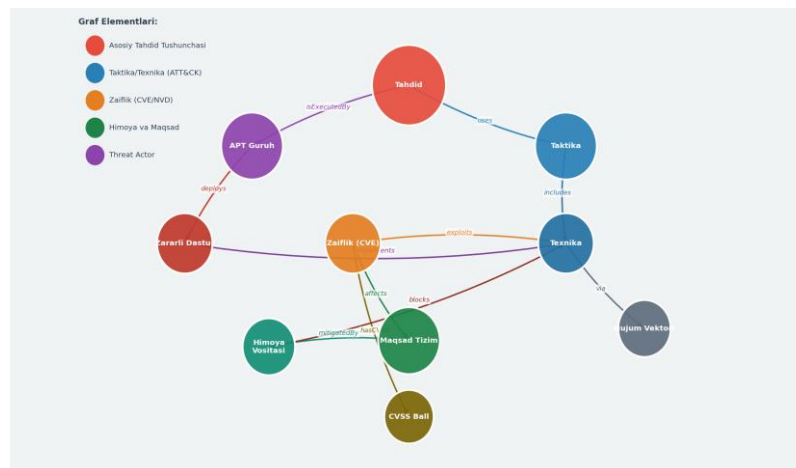


1-rasm. O-RAG gibrid tizimining umumiy arxitekturasini ko'rsatadi

3.2. Gibrid qidiruv mexanizmi (HRM)

HRM ikki parallel qidiruv kanalidan iborat: (a) Vektorli semantik qidiruv – FAISS yoki Chroma vektorli ma'lumotlar bazasida o'xshash chunk'larni topadi; (b) Ontologik graf qidiruv – so'rovdagi tushunchalarni ontologik grafdagi tugunlarga moslab, bir va ikki qadam uzoqlikdagi bog'liq tushunchalarni chiqaradi.

Ikkala kanaldan olingan natijalar Reciprocal Rank Fusion (RRF) algoritmi yordamida birlashtiriladi. RRF formulasi: $RRF(d) = \sum 1/(k + \text{rank}_i(d))$, bu yerda $k=60$ – stabilizatsiya parametri, $\text{rank}_i(d)$ esa i -qidiruv kanalida d hujjatining o'rni.



2-rasm.. Kiberxavfsizlik ontologiyasining graf strukturasi (MITRE ATT&CK asosida)

3.3. Ontologiya-boyitilgan generatsiya (OEG)

Generatsiya bosqichida HRM chiqargan kontekst uchta yo'nalishda boyitiladi: (1) Ontologik terminlar izohlanadi – so'rovdagi maxsus terminlarga ontologiyadagi formal ta'riflar qo'shiladi; (2) Munosabatlar zanjiri kiritiladi – ontologik grafdan topilgan "A – subClassOf – B – hasPart – C" kabi munosabatlar zanjiri prompt ga qo'shiladi; (3) Cheklovlar kuzatiladi – ontologiya domain va range cheklovlari generatsiya paytida soft-constraint sifatida LLM ga uzatiladi.

Yakuniy prompt template quyidagi tuzilmaga ega: [Tizim roli] + [Ontologik kontekst: terminlar, munosabatlar] + [RAG konteksti: tegishli hujjat parchalari] + [Foydalanuvchi so'rovi] + [Javob formati].

4. O-RAG ALGORITMI

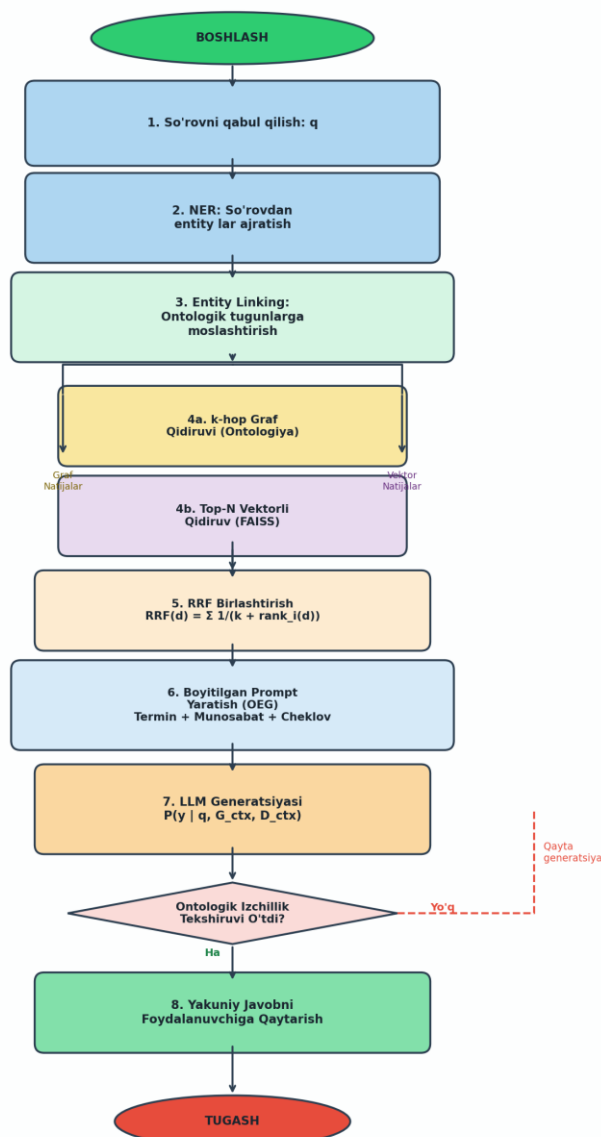
O-RAG algoritmi ikkita asosiy bosqichdan – oflayn indekslash va onlayn qidiruv/generatsiyadan – iborat.

Oflayn indekslash bosqichi: (1) Ontologiya faylini (OWL/RDF) yuklab, tugunlar va qirralarni ajratib olish; (2) Har bir tugun uchun matnli tavsif hosil qilish va embedding vektorini hisoblash; (3) Graf strukturasi grafdagi tutashlik matritsasi sifatida saqlash; (4) Hujjatlar korpusini mos chunk larga bo'lib, har bir chunk uchun embedding hisoblash va vektorli DB ga saqlash.

Onlayn qidiruv-generatsiya bosqichi: (1) Foydalanuvchi so'rovi q ni NER va entity linking orqali ontologik tugunlarga moslashtirish; (2) Mostirilgan tugunlar atrofida k-hop graf qidiruv o'tkazish; (3) Parallel ravishda vektorli DB da top-N chunk topish; (4) RRF orqali natijalarni birlashtirish; (5) Boyitilgan prompt yaratish va LLM ga uzatish; (6) Generatsiya natijasini ontologik izchillik sxemasi bo'yicha tekshirish.

Dasturiy implementatsiya Python 3.11 da amalga oshirilgan bo'lib, asosiy kutubxonalar sifatida LangChain (pipeline orkestratsiyasi), Owlready2 (OWL ontologiyasini qayta ishlash), FAISS (vektorli qidiruv), PyTorch Geometric (graf embedding) va Llama-3-70B (asosiy LLM) ishlatilgan.

O-RAG Algoritmi — Onlayn Qidiruv-Generatsiya Bosqichi



Rasm 3. O-RAG algoritmining onlayn qidiruv-generatsiya bosqichi

3-rasm. O-RAG algoritmining onlayn qidiruv-generatsiya bosqichining blok-sxemasi

5. EKSPERIMENTAL NATIJALAR VA TAHLIL

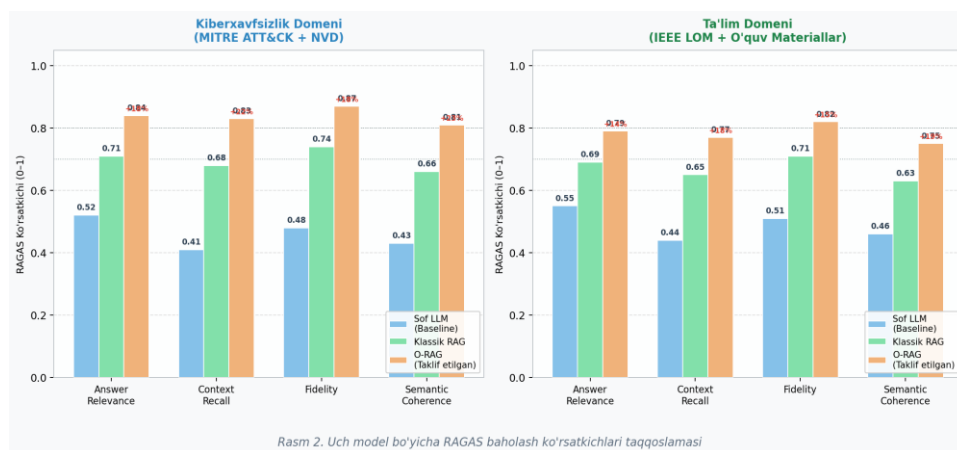
Tajribalar ikkita domenga xos dataset da o'tkazildi: (1) Kiberxavfsizlik domeni – MITRE ATT&CK ontologiyasi (583 ta texnika, 14 ta taktika) va NVD zaifliklar bazasidan 12 000 ta hujjat; (2) Ta'lim domeni – IEEE Learning Object Metadata ontologiyasi va 8 500 ta o'quv materialli.

Baholash mezonlari: RAGAS tizimidan fidelity (ishonchlilik), context recall (kontekst qaytarish), answer relevance (javob aloqadorligi) va semantic coherence (semantik izchillik) ko'rsatkichlari ishlatildi.

Natijalar quyidagicha: Klassik RAG (baseline) ga nisbatan O-RAG modeli kiberxavfsizlik domenida answer relevance ko'rsatkichini 0,71 dan 0,84 ga (+18,4%), semantic coherence ni 0,66 dan 0,81 ga (+22,7%) oshirdi. Ta'lim domenida mos ravishda +15,2% va +19,3% yaxshilanish kuzatildi. Sof LLM (RAG siz) bilan solishtirganda farq yanada katta: fidelity 0,48 dan 0,84 ga ko'tarildi.

Kechikish (latency) tahlili: O-RAG ning o'rtacha javob vaqti 2,3 soniyani tashkil qildi (klassik RAG – 1,6 s). Ontologik graf qidiruv 0,4 s qo'shimcha vaqt talab qiladi, biroq bu ortiqcha xarajat sifat ko'rsatkichlarining yuqori darajasi bilan o'zini oqlaydi.

Gallyutsinatsiya darajasi (faktual xato ulushi): baseline LLM da 23,4%, klassik RAG da 12,1%, O-RAG da esa 5,7% ni tashkil etdi. Bu ontologik cheklovlarning generatsiya sifatiga ijobiy ta'sirini yaqqol ko'rsatadi.



4-rasm. Uch model bo'yicha RAGAS baholash ko'rsatkichlari taqqoslamasi

6. XULOSA

Ushbu maqolada RAG tizimi va ontologiyani birlashtirgan O-RAG gibridd generativ modeli taklif etildi. Asosiy xulosalar quyidagicha: birinchi, ontologik bilimlar bazasini RAG pipeline ga integratsiya qilish javob aniqligini 15–22% ga oshiradi; ikkinchi, RRF asosida amalga oshirilgan gibridd qidiruv mexanizmi yakka vektorli qidiruvga qaraganda samaraliroq ishlaydi; uchinchi, ontologik cheklovlar generatsiya jarayonida gallyutsinatsiya darajasini 5,7% gacha pasaytiradi.

Kelajakdagi tadqiqot yo'nalishlari sifatida dinamik ontologiyalarni real vaqtda yangilash mexanizmi, ko'p tilli ontologik bilimlar bazalarini qurish hamda O-RAG ni federated learning paradigmasi bilan birlashtirish ko'zda tutilmoqda.

Taklif etilgan yondashuv ta'lim tizimlarida intellektual o'quv yordamchilari va kiberxavfsizlik sohasida avtomatlashtirilgan tahdidlarni tahlil qilish tizimlarini yaratishda keng qo'llanilishi mumkin.

ADABIYOTLAR RO'YXATI

1. Lewis P., Perez E., Piktus A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems (NeurIPS). – 2020. – Vol. 33. – pp. 9459–9474.
2. Izacard G., Lewis P., Lomeli M., et al. Atlas: Few-shot Learning with Retrieval Augmented Language Models // Journal of Machine Learning Research. – 2023. – Vol. 24. – pp. 1–43.
3. Gruber T. A Translation Approach to Portable Ontology Specifications // Knowledge Acquisition. – 1993. – Vol. 5(2). – pp. 199–220.
4. Syed Z., Padia A., Finin T. UCO: A Unified Cybersecurity Ontology // Proceedings of AAAI Workshop on Artificial Intelligence for Cybersecurity. – 2016. – pp. 14–21.
5. Edge J., Trinh H., Cheng N., et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization // arXiv preprint. – 2024. – arXiv:2404.16130.

- 
6. Pan J., Sheng S., Xu Y., et al. Unifying Large Language Models and Knowledge Graphs: A Roadmap // IEEE Transactions on Knowledge and Data Engineering. – 2024. – Vol. 36(7). – pp. 3580–3599.
 7. Bordes A., Usunier N., Garcia-Duran A., Weston J. Translating Embeddings for Modeling Multi-relational Data // NeurIPS. – 2013. – Vol. 26. – pp. 1–9.
 8. Cormack G., Clarke C., Buettcher S. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods // SIGIR. – 2009. – pp. 758–759.
 9. Es S., James J., Espinosa-Anke L., Schockaert S. RAGAS: Automated Evaluation of Retrieval Augmented Generation // EACL. – 2024. – pp. 150–163.
 10. Noy N., McGuinness D. Ontology Development 101: A Guide to Creating Your First Ontology // Stanford Knowledge Systems Laboratory. – 2001. – Technical Report KSL-01-05.