SUN’IY INTELLEKT TIZIMLARIDA O‘ZBEK TILI ASTRONOMIYA TERMINOLOGIYASINING RAG ARXITEKTURASI VA TERMINOLOGIK MA’LUMOTLAR BAZASI INTEGRATSIYASI

Fotimaxon BAXTIYAROVA
Filologiya fanlari bo‘yicha (PhD) falsafa doktori
O‘zbekiston jurnalistika va ommaviy
kommunikatsiyalar universiteti
Toshkent, O‘zbekiston
E-mail: fbahtiyarova83@gmail.com
ORCID ID: 0009-0000-1010-3634

Annotatsiya. Mazkur maqolada sun’iy intellekt tizimlarida o’zbek tili astronomiya terminologiyasining aniqligini ta’minlash maqsadida Retrieval-Augmented Generation (RAG) arxitekturasi va terminologik ma’lumotlar bazasi integratsiyasi metodikasi ishlab chiqilgan. 1717 birliklik o’zbek astronomiya tezaurusi FAISS vektorial indeksiga joylashtirilgan va terminologik anchor mexanizmi yordamida LLM generatsiyasida terminologik izchillik ta’minlangan. 500 ta test matni asosida o’tkazilgan eksperiment natijalarida RAG integratsiyasi BLEU skorini 31,4 dan 41,2 gacha, terminologik aniqlik skorini (TAS) esa 58,3% dan 89,1% gacha oshirganligi isbotlangan. Taklif etilayotgan tizim mashina tarjimasi, ilmiy ma’lumot qidirish va ta’lim platformalari uchun amaliy tatbiq etish imkoniyatlarini ochib beradi.

Kalit so‘zlar: RAG arxitekturasi, LLM, astronomiya terminologiyasi, FAISS, terminologik aniqlik, BLEU skori, mashina tarjimasi.

Zamonaviy sun'iy intellekt tizimlari, jumladan katta til modellari (LLM), umumiy matnlarni qayta ishlashda yuqori samaradorlik ko'rsatsa-da, tor soha terminologiyasi — ayniqsa o'zbek tilidagi astronomiya terminlari — bilan ishlashda sezilarli xatoliklarga yo'l qo'yadi. Muammo ikki sababdan kelib chiqadi: birinchidan, o'zbek tili astronomiya terminologiyasi LLM o'quv korpuslarida yetarlicha vakillik qilmaydi; ikkinchidan, astronomiya terminlarining ko'p ma'noliligi (masalan, Oy — samoviy jism va oylik vaqt o'lchovi) modelni noto'g'ri kontekstga yo'naltiradi. Bu muammoni hal qilishning eng istiqbolli yo'li — **Retrieval-Augmented Generation**

(RAG) arxitekturasi asosida ixtisoslashtirilgan terminologik ma'lumotlar bazasini LLM tizimiga ulash metodikasini ishlab chiqishdir.

RAG arxitekturasi 2020 yilda Lewis va boshqalar tomonidan taklif qilingan bo'lib, u statik LLM bilimini dinamik ravishda tashqi bilimlar bazasi bilan boyitadi. Astronomiya sohasida bu yondashuv NASA, ESA va xalqaro astronomiya hamjamiyati tomonidan allaqachon sinovdan o'tkazilgan. Biroq o'zbek tili uchun bunday tizim hali mavjud emas — mazkur tadqiqot shu bo'shliqni to'ldirish va terminologik aniqlikni o'lchash uchun BLEU skori hamda maxsus terminologik aniqlik metrikasini taklif etishga qaratilgan.

Taklif etilayotgan tizim uch asosiy komponentdan iborat: **1) Terminologik indeks** — 1717 birliklik o'zbek astronomiya tezaurusi FAISS (Facebook AI Similarity Search) vektorial ma'lumotlar bazasiga joylashtiriladi; har bir termin uchun ta'rif, sinonimlar, inglizcha va ruscha muqobil, semantik toifa va UAT identifikatori saqlanadi. **2) Retriever moduli** — foydalanuvchi so'rovidan termin nomzodlari ajratib olinadi (NER — Named Entity Recognition) va FAISS indeksidan mos yozuvlar k-NN algoritmi yordamida topiladi (k=5, kosinusli o'xshashlik). **3) Generator moduli** — topilgan terminologik kontekst LLM'ga prompt sifatida uzatiladi, model esa boyitilgan kontekst asosida aniq terminologik tarjima yoki ta'rif yaratadi.

Tizimning asosiy innovatsiyasi — **terminologik anchor mexanizmi**. Oddiy RAG tizimlarida retrieved hujjat va LLM javobi o'rtasida terminologik nomuvofiqlik yuzaga kelishi mumkin. Anchor mexanizmi retrieved terminni hard constraint sifatida belgilaydi: LLM ushbu terminni o'zgartirishi yoki sinonim bilan almashtirishi taqiqlangan. Masalan, "qora tuynuk" termini retrieved bo'lsa, model uni "gravitatsion yig'ilish" deb tarjima qilishi bloklanadi — bu terminologik izchillikning asosiy shartidir. Anchor mexanizmi JSON schema orqali amalga oshirilgan: {"anchor_terms": ["qora tuynuk"], "forbidden_substitutes": ["gravitatsion yig'ilish", "qora o'ra"]}.

Eksperiment uchun 500 ta astronomiya matnidan iborat test to'plami tayyorlangan: 250 ta inglizcha maqola abstrakti (arXiv astro-ph bo'limi) va 250 ta ruscha astronomiya darslik parcha. Tarjima sifati ikki metrika orqali baholangan: **BLEU skori** (matn darajasida umumiy tarjima sifati) va **TAS — Terminologik Aniqlik Skori** (soha terminlarining to'g'ri tarjima qilinish foizi).

Natijalar jadval ko'rinishida: Bazaviy GPT-4 (RAG siz): BLEU = 31.4, TAS = 58.3%. RAG + tezaurus (anchor siz): BLEU = 38.7, TAS = 74.6%. RAG + tezaurus + anchor mexanizmi: BLEU = 41.2, TAS = 89.1%. Ushbu natijalar RAG

integratsiyasi BLEU skorini 31% ga (31.4 → 41.2), terminologik aniqlikni esa 53% ga (58.3% → 89.1%) oshirganini ko'rsatadi. Ayniqsa qo'shma terminlar (ikki va undan ko'p so'zli: qora tuynuk, neytron yulduz, gravitatsion to'lqin) uchun aniqlik 44.1% dan 91.3% gacha oshgan — bu eng muhim natijadir, chunki aynan qo'shma terminlar LLM uchun eng muammoli kategoriya hisoblanadi.

Xatolar tahlili shuni ko'rsatdiki, qolgan 10.9% xatoning asosiy sabablari: neologizmlar (yangi kashfiyot nomlari, masalan Webb teleskopi kashf etgan galaktikalar nomi) — 5.2%; kontekstga bog'liq ko'p ma'nolilik — 3.1%; transliteratsiya nomuvofiqliги (ayrim terminlar o'zbek tilida hali standartlashmagan) — 2.6%. Bu xatolar kelgusida ontologik modellashtirish (1-tezis) bilan integratsiya orqali qisman bartaraf etilishi mumkin.

Taklif etilayotgan RAG + tezaurus + anchor tizimi bir necha amaliy yo'nalishda qo'llanilishi mumkin: **a) Mashina tarjimasi** — xalqaro astronomiya maqolalarini o'zbek tiliga avtomatik tarjima qilish (arXiv, NASA ADS, ESA portali); **b) Ilmiy ma'lumot qidirish** — o'zbek tilidagi so'rov asosida xalqaro ilmiy bazalardan mos materiallarni topish; **c) Ta'lim platformalari** — o'zbek tilida astronomiya kurslarini interaktiv qo'llab-quvvatlash. Tizim modulli tuzilmaga ega bo'lganligi sababli tezaurusni yangilash (yangi terminlar qo'shish) va boshqa tillarga kengaytirish nisbatan osondir.

Kelgusida tadqiqotning asosiy yo'nalishlari: real-time yangilanuvchi tezaurus (yangi astronomiya kashfiyotlari darhol kiritilishi); ontologik modellashtirish bilan to'liq integratsiya (OWL/RDF qatlamini RAG indeksiga qo'shish); va o'zbek astronomiya matnlarining keng annotatsiyalangan korpusini yaratish. Korpus mavjud bo'lsa, TAS metrikasini yanada ishonchli baholash va tizimni fine-tuning orqali takomillashtirish mumkin bo'ladi.

O'zbek tili astronomiya terminologiyasini sun'iy intellekt tizimlarida RAG arxitekturasi va terminologik anchor mexanizmi yordamida boyitish terminologik aniqlikni 58.3% dan 89.1% gacha oshirish imkonini beradi. Ushbu natija o'zbek tilini xalqaro ilmiy AI infratuzilmasiga integratsiyalashtirishning texnik yechimini ko'rsatib, kelajakdagi ilmiy tarjima platformalari, ma'lumot qidirish tizimlari va ta'lim ilovalari uchun poydevor vazifasini o'taydi.

Foydalanilgan adabiyotlar:

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. — P. 9459–9474.

2. Johnson J., Douze M., Jegou H. Billion-scale similarity search with GPUs // IEEE Transactions on Big Data. — 2021. — Vol. 7, № 3. — P. 535–547.

3. Papineni K., Roukos S., Ward T., Zhu W. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting of ACL. — 2002. — P. 311–318.

4. Baxtiyarova F.S. O'zbek tili astronomiya terminologiyasining konseptual-semantik tizimi: filologiya fanlari bo'yicha falsafa doktori (PhD) dissertatsiyasi. — Toshkent: O'zbekiston Milliy universiteti, 2026. — 178 b.