OLIY TA’LIM MUASSASALARIDA GETEROGEN MA’LUMOTLARNI SUN’IY INTELLEKT YORDAMIDA TAHLIL QILISH:

MATEMATIK MODEL VA MAHALLIY YONDASHUV

Jahongir Boltayev,
tayanch doktorant.

Annotatsiya.

Mazkur maqolada O‘zbekiston oliy ta’lim muassasalarida (OTM) HEMIS axborot tizimi, elektron hujjat aylanish platformalari, o‘qitishni boshqarish tizimlari (LMS — Learning Management System) hamda video- va audio-xizmatlar orqali to‘planayotgan geterogen ma’lumotlarni tahlil qilishning matematik modeli ishlab chiqildi. Ma’lumotlar manbalari uchlik orqali formallashtirilib, integratsiya operatori va sun’iy intellekt (Artificial Intelligence, AI) modellarining samaradorlik mezonlari aniq formulalar bilan keltirildi. Mashinaviy o‘qitish (Machine Learning) modellarining qiyosiy tahlili o‘tkazildi. Shuningdek, ma’lumotlar suvereniteti va shaxsiy ma’lumotlar himoyasi sababli global katta til modellaridan (Large Language Models, LLM) bevosita foydalanishning huquqiy va texnik cheklovlari hamda mahalliy yondashuvning afzalliklari asoslandi.

Kalit so‘zlar:

geterogen ma’lumotlar, sun’iy intellekt, HEMIS, oliy ta’lim, mashinaviy o‘qitish, katta til modellari, ma’lumotlar suvereniteti, ma’lumotlar integratsiyasi.

Аннотация.

В статье разработана математическая модель анализа гетерогенных данных, накапливаемых в высших учебных заведениях (ВУЗ) Узбекистана через информационную систему HEMIS, платформы электронного документооборота, системы управления обучением (LMS), видео- и аудиосервисы. Источники данных формализованы через тройку признаков, представлен оператор интеграции и приведены формальные метрики оценки моделей искусственного интеллекта. Проведён сравнительный анализ моделей машинного обучения. Также обоснованы правовые и технические ограничения прямого использования глобальных больших языковых моделей (LLM) в связи с суверенитетом данных и защитой персональных данных, и обоснованы преимущества локального подхода.

Ключевые слова:

гетерогенные данные, искусственный интеллект, HEMIS, высшее образование, машинное обучение, большие языковые модели, суверенитет данных, интеграция данных.

Abstract.

This article develops a mathematical model for the analysis of geterogeneous data accumulated in Uzbek higher education institutions (HEIs) through the HEMIS information system, electronic document management platforms, Learning Management Systems (LMS), and video and audio services. Data sources are formalized through a triple of features, an integration operator is introduced, and formal evaluation metrics for artificial intelligence (AI) models are presented. A comparative analysis of machine learning models is performed. The article also substantiates the legal and technical limitations of using global large language models (LLMs) due to data sovereignty and personal data protection concerns, and justifies the advantages of a local approach.

Key words:

geterogeneous data, artificial intelligence, HEMIS, higher education, machine learning, large language models, data sovereignty, data integration.

Kirish.

O‘zbekiston Respublikasi Prezidentining 2020-yil 5-oktabrdagi PF-6079-son farmoni bilan tasdiqlangan “Raqamli O‘zbekiston — 2030” strategiyasi[1] va 2019-yil 8-oktabrdagi PF-5847-son farmoni bilan tasdiqlangan oliy ta’limni 2030-yilgacha rivojlantirish konsepsiyasi[2] doirasida oliy ta’lim muassasalarining (OTM) raqamli infratuzilmasi tubdan yangilanmoqda. Davlat OTMlariga keng joriy etilgan HEMIS yagona axborot tizimi[3] talabalar kontingenti, akademik faoliyat va o‘quv jarayoniga oid katta hajmdagi ma’lumotlarni elektron tarzda yig‘ish hamda boshqarish imkonini beradi.

Shu bilan birga, OTMlarda HEMISdan tashqari turli tabiatdagi axborot platformalari parallel ravishda faoliyat yuritmoqda. Bularga LMS (Learning Management System — Moodle, Open edX kabilar) — kurs materiallari va talabalar faolligi loglari; elektron hujjat aylanish tizimlari (electronic document management, E-XAT va o‘xshashlari) — administrativ hujjatlar; kutubxona axborot tizimlari — bibliografik va o‘qish faoliyati yozuvlari; video- va audio-xizmatlar — masofaviy darslar yozuvlari va lingafon mashqlari; korporativ pochta va xabar almashish vositalari kiradi.

Bu manbalardan kelayotgan ma’lumotlar tabiatan geterogendir (geterogeneous): ular bir-biridan format (raqamli, matnli, fayl, video, audio), tuzilish (strukturalashgan, strukturalashmagan, yarim-strukturalashgan) va semantika jihatidan keskin farqlanadi. An’anaviy hisobot tizimlari bu xilma-xil ma’lumotlarni yagona tahlil maydonida ko‘rib chiqishga qodir emas.

1 O‘zbekiston Res
farmoni // Qonunchilik ma’lumotlari milliy bazasi, 06.10.2020.
2 O‘zbeki
rivojlantirish konsepsiyasini tasdiqlash to‘g‘risida”gi farmoni.
3 HEMIS (Hi
davlat oliy ta’lim muassasalariga joriy etilgan yagona axborot tizimi. Rasmiy portal: hemis.uz

Tadqiqotning maqsadi — OTMlardagi geterogen ma'lumotlar oqimini matematik jihatdan formallashtirish, sun'iy intellekt (Artificial Intelligence, AI) usullari yordamida tahlil qilish modelini ishlab chiqish hamda global katta til modellaridan (Large Language Models, LLM) bevosita foydalanish mumkin bo'lmagan sharoitda mahalliy yondashuvni asoslashdir.

Vazifalar: (a) OTMdagi ma'lumot manbalarini matematik formallashtirish; (b) integratsiya (integration) operatori va tahlil modelini ishlab chiqish; (v) mashinaviy o'qitish (Machine Learning, ML) modellarining qiyosiy samaradorligini baholash; (g) global LLMlardan foydalanish cheklovlarini huquqiy va texnik jihatdan tahlil qilish; (d) mahalliy yondashuv arxitekturasini taklif etish.

Mavzuga oid adabiyotlarning tahlili

Ma`lumotlar integratsiyasining nazariy asoslari M. Lenzerini tomonidan ishlab chiqilgan bo'lib, muallif manba va global sxemalar o'rtasidagi mosliklarni (Global-as-View, Local-as-View, Global-Local-as-View — GAV, LAV, GLAV yondashuvlari) formal mantiq tilida ifoda etgan[1]. A. Doan, A. Halevy va Z. Ives klassik monografiyasida integratsiyaning federativ (federated), virtual va materializatsiyalangan (materialized) yondashuvlari tizimli yoritilgan[2].

Ta'lim sohasida ma'lumotlar tahlili mustaqil ilmiy yo'nalish — ta'lim analitikasi (Learning Analytics) sifatida G. Siemens va P. Long tomonidan asoslangan[3]. Ta`limda ma`lumotlar tog'-konchiligi (Educational Data Mining, EDM) yo'nalishini C. Romero va S. Ventura rivojlantirib, 1995–2019 yillardagi tadqiqotlarni umumlashtirgan[4].

Mashinaviy o'qitishning ansambl usullaridan (ensemble methods) tasodifiy o'rmon (Random Forest) L. Breiman tomonidan[5], gradient bo'sting (Gradient Boosting, XGBoost) esa T. Chen va C. Guestrin tomonidan ishlab chiqilgan[6]. Chuqur o'qitishning (Deep Learning) fundamental asoslari Y. LeCun, Y. Bengio va G. Hinton ishida bayon etilgan[7].

Katta til modellari (LLM) bo'yicha asosiy nazariy ish T. Brown va boshqalar tomonidan nashr etilgan[8]; unda GPT-3 modelining kam misolli o'qitish (few-shot learning) qobiliyati namoyish etilgan. LLM larning imkoniyatlari va xavf-xatarlari R. Bommasani va boshqalar tomonidan tizimli o'rganib chiqilgan[9]. Faktik xato (hallucination) muammosini Z. Ji va boshqalar tahlil qilishgan[10]; mualliflar hallyusinatsiyani LLM larning intrinsik xususiyati sifatida ko'rsatishgan.

O'zbekistonda shaxsiy ma'lumotlarni qayta ishlash O'RQ-547-son qonun[11], Yevropa Ittifoqida esa GDPR me'yorlari[12] bilan tartibga solinadi. Kam resursli tillarda (low-resource languages) LLM lar samaradorligining pasayishi K. Ahuja va boshqalarning MEGA benchmarki orqali asoslangan[13].

Adabiyotlar tahlili shuni ko'rsatadiki, OTM darajasidagi geterogen ma'lumotlarni mahalliy huquqiy va texnik sharoitlarda matematik formallashtiruvchi yaxlit yondashuv yetarlicha tadqiq etilmagan. Aynan shu bo'shliq mazkur tadqiqotning ilmiy yangiligini belgilaydi.

Tadqiqot metodologiyasi va matematik model

Tadqiqot metodologiyasi pragmatik falsafiy yondashuvga asoslanib, formal-matematik va eksperimental usullarni birlashtiradi. Quyida geterogen ma'lumotlarni va ularni tahlil qilishning matematik formalizatsiyasi keltiriladi.

Ta'rif 1. Bo'lsin

$$D = \{ D_1, D_2, \ldots, D_n \}, \qquad (1)$$

OTMdagi ma'lumot manbalari to'plami. Har bir D_i manba uchligi orqali tavsiflanadi:

$$D_i = \langle S_i, F_i, \sigma_i \rangle, \qquad (2)$$

bu yerda S_i — manbaning sxemasi (schema), F_i — formati (format), σ_i — semantik anglatma (semantic mapping).

1 Lenzerini M. Data Integration: A Theoretical Perspective // Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS '02). — ACM, 2002. — P. 233–246.
2 Doan A., Halevy A., Ives Z. Principles of Data Integration. — Morgan Kaufmann, 2012. — 520 p.
3 Siemens G., Long P. Penetrating the Fog: Analytics in Learning and Education // EDUCAUSE Review. — 2011. — Vol. 46, No. 5. — P. 30–40.
4 Romero C., Ventura S. Educational data mining and learning analytics: An updated survey // WIREs Data Mining and Knowledge Discovery. — 2020. — Vol. 10, No. 3. — e1355.
5 Breiman L. Random Forests // Machine Learning. — 2001. — Vol. 45, No. 1. — P. 5–32.
6 Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). — ACM, 2016. — P. 785–794.
7 LeCun Y., Bengio Y., Hinton G. Deep Learning // Nature. — 2015. — Vol. 521. — P. 436–444.
8 Brown T., Mann B., Ryder N. et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems (NeurIPS 2020). — 2020. — Vol. 33. — P. 1877–1901.
9 Bommasani R., Hudson D.A., Adeli E. et al. On the Opportunities and Risks of Foundation Models. — Stanford CRFM Technical Report, 2021. — arXiv:2108.07258.
10 Ji Z., Lee N., Frieske R. et al. Survey of Hallucination in Natural Language Generation // ACM Computing Surveys. — 2023. — Vol. 55, No. 12. — P. 1–38.
11 O'zbekiston Respublikasining 2019-yil 2-iyuldagi O'RQ-547-son "Shaxsga doir ma'lumotlar to'g'risida"gi qonuni.
12 General Data Protection Regulation (EU) 2016/679. — Official Journal of the European Union, L 119, 4.5.2016.
13 Ahuja K., Diddee H., Hada R. et al. MEGA: Multilingual Evaluation of Generative AI // Proceedings of EMNLP 2023. — Association for Computational Linguistics, 2023.

Ta’rif 2 (Format to‘plami). Format chekli to‘plami:

$$F = \{f_{num}, f_{txt}, f_{doc}, f_{vid}, f_{aud}, f_{str}\}, \quad (3)$$

bu yerda f_{num} — raqamli (numeric, HEMIS reytinglari, davomat foizi); f_{txt} — matnli (textual, talabalar javoblari, so‘rovnomalar); f_{doc} — fayl/hujjat (document, PDF, DOCX, hujjat aylanish tizimlaridan); f_{vid} — video (onlayn darslar); f_{aud} — audio (lingafon, podkast); f_{str} — yarim-strukturalashgan (semi-structured: JSON, XML, LMS loglari).

Ta’rif 3 (Geterogenlik).

Manbalar to‘plami D geterogen deyiladi, agar

$$\exists\, i, j \in \{1, \ldots, n\},\ i \neq j:\ (S_i \neq S_j) \vee (F_i \neq F_j) \vee (\sigma_i \neq \sigma_j). \quad (4)$$

Ta’rif 4 (Integratsiya operatori).

I operator manbalarni global sxemaga aks ettiradi:

$$I : D_1 \times D_2 \times \ldots \times D_n \to D_G, \quad (5)$$

bu yerda D_G — birlashtirilgan global ma’lumotlar to‘plami. I operatorni ETL (Extract, Transform, Load) bosqichlari kompozitsiyasi sifatida ifoda etamiz:

$$I = L \circ T \circ E, \quad (6)$$

ya’ni E (Extraction) — manbalardan ajratib olish, T (Transformation) — formatlar va sxemalarni unifikatsiyalash, L (Loading) — omborga yuklash.

Ta’rif 5 (Tasniflash modeli). Bo‘lsin

$$X \in \mathbb{R}^d, \quad y \in \{0, 1\}, \quad (7)$$

bu yerda X — talabaning d ta xususiyati vektori (HEMIS bahosi, davomat foizi, LMS faolligi, kutubxona faolligi, so‘rovnoma indekslari va h.k.), y — akademik muvaffaqiyat binari. Tasniflash (classification) modeli f bashorat funksiyasini aniqlaydi:

$$\hat{y} = f(X;\, \theta), \quad (8)$$

bu yerda θ — model parametrlari.

Gradient bo‘sting (XGBoost) modelining maqsad funksiyasi quyidagicha[14]:

$$L(\theta) = \Sigma_i\, \ell(y_i, \hat{y}_i) + \Sigma_k\, \Omega(f_k), \quad (9)$$

$$\Omega(f) = \gamma\, T + \tfrac{1}{2}\, \lambda\, \|w\|^2, \quad (10)$$

bu yerda ℓ — yo‘qotish funksiyasi (loss function), T — daraxt yaproqlari soni, w — yaproq og‘irliklari (leaf weights) vektori, γ, λ — regulyarizatsiya (regularization) parametrlari.

Ta’rif 6 (Samaradorlik metrikalari).

TP, TN, FP, FN — chalkashlik matritsasining (confusion matrix) mos elementlari deylik. U holda:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN), \quad (11)$$

$$Precision = TP / (TP + FP), \quad (12)$$

$$Recall = TP / (TP + FN), \quad (13)$$

$$F1 = 2 \cdot Precision \cdot Recall / (Precision + Recall). \quad (14)$$

k-bo‘lakli o‘zaro tekshiruv (k-fold cross-validation) metrikasi:

$$CV_k = (1 / k) \cdot \Sigma_{j=1}^{\ldots k}\, M(f_{(j)}), \quad (15)$$

bu yerda $f_{(j)}$ — j-bo‘lak tasdiqlash to‘plami sifatida olib o‘qitilgan model, M — (11)–(14) dan biri.

Eksperimentda 5-fold cross-validation qo‘llanildi; o‘rgatish/tasdiqlash/sinov nisbati (training/validation/test split) 70/15/15 ga teng. Gipermetrlar (hyperparameters) grid search orqali optimallashtirildi. Hisob-kitoblar Python 3.11, scikit-learn 1.4, XGBoost 2.0, PyTorch 2.2 muhitida o‘tkazildi.

OTMdagi geterogen ma’lumot manbalarining tasnifi

(2)–(3) formulalarga muvofiq, OTMda eng tez-tez uchraydigan manbalar 1-jadvalda umumlashtirilgan.

14 Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16). — ACM, 2016. — P. 785–794. — Equation 1, P. 786.

1-jadval.

OTMdagi geterogen ma'lumot manbalari (formula (2) bo'yicha)

Platforma / manba	Ma'lumot turi (F)	Misol va format	Struktura (S)
HEMIS	Raqamli (f_num)	Reyting, kreditlar, davomat foizi (SQL/REST API)	Strukturalashgan (structured)
Moodle / LMS	Aralash (f_str, f_txt)	Faollik loglari, forum yozuvlari (JSON, XML)	Yarim-strukturalashgan (semi-structured)
Elektron hujjat aylanish (E-XAT)	Fayl (f_doc)	Buyruqlar, hisobotlar (PDF, DOCX)	Strukturalashmagan (unstructured)
Video xizmati	Video (f_vid)	Onlayn dars yozuvlari (MP4, WebM) + meta-ma'lumotlar	Strukturalashmagan
Lingafon kabineti	Audio (f_aud)	Talaffuz mashqlari (MP3, WAV)	Strukturalashmagan
Kutubxona ATI	Matnli + raqamli	Bibliografik yozuvlar, o'qish tarixi (MARC, SQL)	Strukturalashgan
Talabalar so'rovnomalari	Matnli (f_txt)	Erkin matnli javoblar (CSV, matn)	Aralash
Korporativ pochta	Matnli + fayl	Xat-xabarlar, ilovalar (IMAP)	Yarim-strukturalashgan

Manba: muallif tomonidan tizimlashtirildi.

HEMIS yadrosida raqamli ma'lumotlar ((3) formuladagi f_{num}) to'planadi: talabalar reytingi, kreditlar tizimi (ECTS — European Credit Transfer System bo'yicha), davomat foizi, oraliq va yakuniy nazorat ballari, bitiruv reytingi. Ushbu ma'lumotlar relyatsion bazada (relational database) saqlanib, REST API orqali tashqi tizimlarga ulanish imkoniyatiga ega.

Moodle va shunga o'xshash LMSlar talabalar faolligi loglarini JSON/XML formatida saqlaydi (f_{str}): kursga kirish vaqti, materiallarni ko'rish soni, testlardagi javob vaqti, forumdagi xabarlar. Bu yarim-strukturalashgan ma'lumotlar talabaning o'rganish naqshini (learning pattern) aks ettiradi.

Elektron hujjat aylanish platformalari (electronic document management — E-XAT va analoglari) administrativ jarayonlarni (buyruqlar, hisobotlar, ma'lumotnomalar) PDF/DOCX formatida saqlaydi (f_{doc}). Video xizmatlari (Zoom yozuvlari, ichki video-arxivlar) f_{vid} turidagi katta hajmli fayllar va metama'lumotlar (vaqt, ishtirokchilar, faollik) bilan ishlaydi. Lingafon kabinetlari va talaffuz mashqlari f_{aud} formatdagi audio-fayllarni hosil qiladi.

Eksperimental tahlil va natijalar

(8)–(10) bo'yicha aniqlangan modellar talabalar akademik muvaffaqiyatini bashorat qilish vazifasida sinovdan o'tkazildi. Kirish o'zgaruvchilari (X) sifatida HEMISdagi raqamli ko'rsatkichlar, Moodle log-loridan olingan faollik indekslari, kutubxona statistikalari va so'rovnoma natijalari qo'llanildi. Sinov ma'lumotlar to'plami n = 3 856 talabaning to'rt semestrlik anonimlashtirilgan (anonymized) ma'lumotlarini o'z ichiga oldi[1]. (11)–(14) metrikalari bo'yicha natijalar 2-jadvalda keltirilgan.

1 Ushbu ra

2-jadval.

(11)–(14) formulalar bo‘yicha modellarning qiyosiy samaradorligi

Model	Accuracy	Precision	Recall	F1-score
Logistik regressiya (Logistic Regression)	0,74	0,72	0,71	0,71
Tasodifiy o’rmon (Random Forest)	0,82	0,80	0,79	0,79
XGBoost (Gradient Boosting)	**0,87**	**0,86**	**0,84**	**0,85**
DNN (Deep Neural Network, 4 qatlam)	0,85	0,83	0,82	0,83

Manba: muallif tomonidan o’tkazilgan eksperimentlar (5-fold cross-validation).

Natijalarga ko‘ra, gradient bo‘sting (XGBoost) modeli barcha asosiy metrikalar bo‘yicha eng yuqori qiymatlarni ko‘rsatdi: Accuracy = 0,87, F1 = 0,85. Bu xususiyatlar fazosi (feature space) yuqori o‘lchamlik va geterogen tipdagi xususiyatlarga ega bo‘lgan vaziyatlarda XGBoost ning ansambl tabiati va regulyarizatsiya ((10) formula) afzalligi bilan izohlanadi[2].

Chuqur neyron tarmoq (Deep Neural Network, DNN) modelining natijasi (F1 = 0,83) XGBoostga yaqin bo‘ldi, biroq o‘qitish vaqti taxminan 8 marta uzoq va hisoblash resurslari talabi yuqori. Bu o‘rta hajmdagi ($n \leq 10^4$) ma’lumotlar uchun ansambl daraxtlari (tree ensembles) ko‘pincha chuqur neyron tarmoqlardan kam emasligini ko‘rsatadi[3].

Integratsiya operatori (5) ning amaliy ta’siri quyidagicha baholandi: faqat HEMIS ma’lumotlari asosida F1 = 0,71; HEMIS + LMS qo’shilganda F1 = 0,79; HEMIS + LMS + kutubxona to‘plamida F1 = 0,85. Bu integratsiyaning marjinal foydasi har bir manba qo‘shilishi bilan kamayuvchi (diminishing returns) xarakterga ega ekanligini, ammo qo‘shimcha manba samaradorlikni izchil oshirishini ko‘rsatadi.

Global katta til modellaridan foydalanish cheklovlari

Bugungi kunda ChatGPT, Claude, Gemini kabi global katta til modellari (LLM) keng tarqalgan. Biroq, ularni O‘zbekiston OTMlarining geterogen ma’lumotlarini tahlil qilishda bevosita qo‘llashning bir qator jiddiy huquqiy, texnik va etik cheklovlari mavjud.

1. Ma’lumotlar suvereniteti va huquqiy cheklovlar. O‘RQ-547-son qonun[4] shaxsiy ma’lumotlarni qayta ishlash uchun ularning Respublika hududidagi serverlarda saqlanishini va davlat hisobida bo‘lishini belgilaydi. Global LLM provayderlari (OpenAI, Anthropic, Google) so‘rovlarni AQSh va Yevropa serverlarida qayta ishlaydi, bu esa qonun talablariga zid keladi. GDPR me’yorlari[5] ham ma’lumotlarni uchinchi davlatlarga uzatishni cheklaydi.

2. Til qo‘llab-quvvatlashning zaifligi. K. Ahuja va boshqalarning MEGA benchmarki bo‘yicha[6], LLM larning kam resursli tillardagi (low-resource languages) samaradorligi ingliz tiliga nisbatan 20–40 % gacha past bo‘lishi mumkin. O‘zbek tili (xususan lotin alifbosida) ham shu toifaga kiradi, bu esa LLM larning mahalliy matnli ma’lumotlar bilan ishlashini cheklaydi.

2 Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD ’16). — ACM, 2016. — P. 785–794. — P. 791-792.
3 LeCun Y., Bengio Y., Hinton G. Deep Learning // Nature. — 2015. — Vol. 521. — P. 436–444. — P. 438.
4 O‘zbekiston Respublikasining 2019-yil 2-iyuldagi O‘RQ-547-son “Shaxsga doir ma’lumotlar to‘g‘risida”gi qonuni.
5 General Data Protection Regulation (EU) 2016/679. — Official Journal of the European Union, L 119, 4.5.2016.
6 Ahuja K., Diddee H., Hada R. et al. MEGA: Multilingual Evaluation of Generative AI // Proceedings of EMNLP 2023. — Association for Computational Linguistics, 2023.

3. Hallyusinatsiya (hallucination) muammosi. Z. Ji va boshqalar[1] hallyusinatsiyani LLM larning intrinsik xususiyati sifatida koʻrsatishgan: model fakt jihatidan notoʻgʻri, lekin tashqi tomondan ishonchli maʼlumotni generatsiya qilishi mumkin. Akademik qarorlar (baho bashorati, ilmiy ish baholash, stipendiya tavsiyalari) kabi masʼuliyatli sohalarda bunday xato narxi yuqori.

4. Mahalliy kontekstga moslashmaganlik. Global LLM larda HEMIS sxemasi, Oʻzbekiston OTMlarining ichki nizomlari, mahalliy taʼlim standartlari haqida bilim yetishmaydi. Model soʻrovga javoban tashqi domendagi sxemalarni ekstrapolyatsiya qilishi (extrapolation) mumkin.

5. Iqtisodiy va texnologik bogʻliqlik. API toʻlovlari xorijiy valyutada amalga oshiriladi. Geosiyosiy oʻzgarishlar yoki sanksiyalar xizmatdan foydalanishni bir tomonlama cheklashi mumkin. Asos modellar (foundation models) markazlashishi bilan bogʻliq sistemli risklar[2] alohida muhokama qilingan.

6. Takrorlanuvchanlik (reproducibility) muammosi. LLM larning chiqishi bir xil kirish uchun ham har xil boʻlishi mumkin (temperatura > 0), modellar provayder tomonidan ogohlantirishsiz oʻzgartiriladi. Bu ilmiy tadqiqotning takrolanuvchanlik prinsipiga zid.

7. Shaffoflik va auditga yopiqlik. Yopiq LLM larning ichki ishlashi mexanizmlarini tekshirib boʻlmaydi (black-box). Bu taʼlimda qabul qilinadigan qarorlarning shaffofligi va himoya qilinishini cheklaydi.

Cheklovlarning oqibatlari va mahalliy yondashuv

Ushbu cheklovlarning amaliy oqibatlari quyidagilarda namoyon boʻladi: (a) shaxsiy maʼlumotlarni global LLMga uzatish huquqiy javobgarlikka olib kelishi mumkin; (b) oʻzbek tilidagi tahlillarning sifati past boʻladi; (v) tadqiqot natijalari takrorlanmaydi; (g) OTM xizmatlari xorijiy provayderlarga bogʻliq boʻlib qoladi.

Ushbu oqibatlarni yumshatish uchun quyidagi mahalliy yondashuv tavsiya etiladi:

(a) Ochiq vaznli (open-weight) modellarni mahalliy serverda joylashtirish (on-premise deployment). Meta LLaMA-2/3[3], Mistral, Qwen kabi ochiq vaznli modellarni respublika hududidagi infratuzilmada joylashtirish maʼlumotlarning tashqariga chiqishini istisno qiladi va OʻRQ-547 talablariga muvofiq keladi.

(b) Oʻzbek tilida nozik sozlash (fine-tuning). Ochiq modellarni HEMIS, Moodle va ichki hujjat aylanish maʼlumotlari asosida tor sohaga moslashtirish til qoʻllab-quvvatlashning zaifligini bartaraf etadi.

(v) Gibrid arxitektura (hybrid architecture). Strukturalashgan maʼlumotlar (HEMIS reytingi, davomat) uchun anʼanaviy ML usullari (XGBoost, Random Forest — (8)–(10) formulalar), strukturalashmagan maʼlumotlar (matn, hujjat) uchun mahalliy LLM ishlatish samaradorlik va xavfsizlik oʻrtasidagi murosani taʼminlaydi.

(g) Federativ oʻqitish (federated learning). B. McMahan va boshqalar tomonidan taklif etilgan federativ yondashuv[4] OTMlarga maʼlumotlarini almashmasdan turib umumiy modelni birgalikda oʻqitishga imkon beradi.

(d) Anonimlashtirish (anonymization) protokollari. Shaxsiy identifikatorlarni xeshlash (hashing), k-anonimlik (k-anonymity, $k \geq 5$) prinsipini joriy etish va differential privacy mexanizmlarini qoʻllash zarur.

(e) Verifikatsiya qatlami (guardrails). LLM natijalarini boshqa manbalar (qoidalar bazasi, statistik testlar) bilan tasdiqlovchi avtomatik tekshiruv tizimlarini joriy etish hallyusinatsiya xavfini kamaytiradi.

Xulosa va takliflar

Mazkur tadqiqotda OTMdagi geterogen maʼlumotlarni tahlil qilishning matematik modeli va mahalliy yondashuvi taklif etildi. Asosiy natijalar quyidagilardan iborat:

1. Geterogen maʼlumotlar manbalari (2) formuladagi uchlik orqali, integratsiya jarayoni esa (5)–(6) formulalarda kompozitsion operator $I = L \circ T \circ E$ sifatida formallashtirildi.

2. (8)–(10) boʻyicha sinovdan oʻtkazilgan toʻrt model ichida XGBoost eng yuqori F1 = 0,85 koʻrsatkichini koʻrsatdi; integratsiya operatorining qoʻllanishi samaradorlikni alohida manbalarga nisbatan 14 punktga oshirdi.

3. Global LLMlardan Oʻzbekiston OTMlarida bevosita foydalanish OʻRQ-547 qonuni, hallyusinatsiya muammosi, til qoʻllab-quvvatlashning zaifligi va texnologik bogʻliqlik sabablari bilan cheklangan.

4. Mahalliy yondashuv — on-premise ochiq modellar, fine-tuning, gibrid arxitektura, federativ oʻqitish va anonimlashtirish protokollari — ushbu cheklovlarni hal qiluvchi yaxlit yechim sifatida asoslandi.

Amaliy tavsiyalar: HEMIS, LMS, hujjat aylanish va kutubxona platformalarini yagona maʼlumotlar omborida (data warehouse) birlashtirish; on-premise ochiq vaznli modellarni joriy etish; OTMlarda maʼlumotlar olimi (data scientist) kadrlari tayyorlash; izohlanuvchi sunʼiy intellekt (Explainable AI, XAI) yondashuvlarini qaror qabul qilish jarayonlarida qoʻllash; shaxsiy maʼlumotlar himoyasi protokollarini standartlashtirish.

Kelajak tadqiqot yoʻnalishlari: oʻzbek tiliga moslashtirilgan ochiq LLM larni ishlab chiqish; multimodal (matn + audio + video) tahlil modellari; OTMlar aro federativ oʻqitish stendi; taʼlim sohasiga moslashtirilgan SI etikasi meʼyorlarini ishlab chiqish.

1 Ji Z., Lee N., Frieske R. et al. Survey of Hallucination in Natural Language Generation // ACM Computing Surveys. — 2023. — Vol. 55, No. 12. — P. 1–38.

2 Bommasani R., Hudson D.A., Adeli E. et al. On the Opportunities and Risks of Foundation Models. — Stanford CRFM Technical Report, 2021. — arXiv:2108.07258.

3 Touvron H., Lavril T., Izacard G. et al. LLaMA: Open and Efficient Foundation Language Models. — arXiv:2302.13971, 2023.

4 McMahan B., Moore E., Ramage D. et al. Communication-Efficient Learning of Deep Networks from Decentralized Data // AISTATS 2017. — PMLR, 2017. — Vol. 54. — P. 1273–1282.