UNIVERSITET MA'LUMOTLAR EKOTIZIMIDA INTELLEKTUAL TAHLIL: HETEROGEN MANBALARNI BIRLASHTIRISHNING MATEMATIK MODELI VA LOKALIZATSIYALANGAN ARXITEKTURA

Boltayev Jahongir Erkin o’g’li
Termiz davlat universiteti
E-mail: boltayev.jahongir@gmail.com
Tel: +998 97 617 96 96

Annotatsiya – *Zamonaviy oliy ta’lim muassasalari (OTM) HEMIS, o‘qitishni boshqarish tizimlari (LMS), elektron hujjat aylanish platformalari hamda video- va audio-xizmatlar tarkibida bir-biridan format, tuzilish va semantik jihatdan tafovutlanuvchi ma’lumotlar oqimini hosil qiladi. Mazkur xilma-xillik an’anaviy hisobot vositalarining tahliliy quvvatini cheklab qo‘yadi. Ushbu tadqiqotda heterogen manbalarni formal tavsiflashga mo‘ljallangan uchlik bazasidagi matematik model, ETL prinsiplariga asoslangan integratsiya operatori va sun’iy intellekt algoritmlarini baholash mezonlari taqdim etilgan. To‘rt model qiyosida XGBoost eng yuqori F1 = 0,85 qiymatini qayd etgan. Shuningdek, O‘zbekiston huquqiy maydonidagi ma’lumotlar suvereniteti talablari hamda texnik omillar sababli xalqaro katta til modellaridan (LLM) bevosita foydalanishning to‘siqlari va lokal implementatsiya — mahalliy serverlardagi ochiq vaznli modellar, nozik sozlash hamda federativ o‘qitish — afzalliklari ko‘rib chiqilgan.*

Kalit so‘zlar: *heterogen ma’lumotlar, sun’iy intellekt, HEMIS, oliy ta’lim, integratsiya operatori, mashinaviy o‘qitish, katta til modellari, ma’lumotlar suvereniteti.*

KIRISH

Raqamli texnologiyalarning oliy ta’lim sohasiga keng kirib kelishi ma’lumotlar oqimi va xilma-xilligining mislsiz darajada o‘sishiga olib kelmoqda. O‘zbekistonda mazkur jarayon Prezidentning 2020-yil 5-oktabrdagi PF-6079-son farmoni bilan tasdiqlangan “Raqamli O‘zbekiston — 2030” strategiyasi[10] va 2019-yil 8-oktabrdagi PF-5847-son farmoni asosida tasdiqlangan oliy ta’lim 2030 konsepsiyasi[11] doirasida amalga oshirilmoqda.

Islohotlar ramkasida HEMIS yagona axborot tizimi[12] deyarli barcha davlat OTMlariga joriy etilib, talabalar kontingenti, akademik faoliyat va o‘quv jarayoniga oid raqamli ko‘rsatkichlarni elektron yig‘ishni ta’minlaydi. Biroq, HEMIS —

[10]O‘zbekiston Respublikasi Prezidentining 2020-yil 5-oktabrdagi PF-6079-son “Raqamli O‘zbekiston — 2030” strategiyasini tasdiqlash to‘g‘risidagi farmoni // Qonunchilik ma’lumotlari milliy bazasi, 06.10.2020.

[11]O‘zbekiston Respublikasi Prezidentining 2019-yil 8-oktabrdagi PF-5847-son “O‘zbekiston Respublikasi oliy ta’lim tizimini 2030-yilgacha rivojlantirish konsepsiyasini tasdiqlash to‘g‘risida”gi farmoni.

[12]HEMIS (Higher Education Management Information System) — O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi tomonidan davlat oliy ta’lim muassasalariga joriy etilgan yagona axborot tizimi. Rasmiy portal: hemis.uz

universitet ma’lumotlar ekotizimining bir qismi xolos.

Universitet ichida HEMIS bilan parallel ravishda quyidagi platformalar funksiyonal: o'qitishni boshqarish tizimlari (LMS — Learning Management System: Moodle, Open edX) kurs kontenti va talabalar faolligi loglarini saqlaydi; elektron hujjat aylanish vositalari (electronic document management — E-XAT va analoglari) administrativ hujjatlarni qayta ishlaydi; kutubxona axborot tizimlari bibliografik va o'qish statistikasini yig'adi; video- hamda audio-xizmatlar masofaviy darslar va lingafon mashqlarini arxivlaydi; korporativ kommunikatsiya tizimlari xat-xabar oqimini boshqaradi.

Manbalardan keluvchi ma'lumotlar tabiatan heterogendir (heterogeneous): raqamli (HEMIS reytinglari), matnli (so'rovnomalar), fayl shaklidagi (hujjatlar), video va audio (darslar), shuningdek yarim-strukturalashgan (LMS loglari). Ushbu manbalarni yagona tahlil maydonida ko'rib chiqish uchun maxsus matematik apparat va sun'iy intellekt vositalari talab etiladi.

Ushbu ishning maqsadi — OTM ma'lumotlar ekotizimini matematik jihatdan formallashtirish, sun'iy intellekt (Artificial Intelligence, AI) usullari bilan tahlil qilish modelini ishlab chiqish hamda xalqaro katta til modellaridan (Large Language Models, LLM) bevosita foydalanish cheklangan sharoitda lokalizatsiyalangan yechimni asoslashdir.

Ish vazifalari: (1) manbalarning matematik tavsiflash apparatini taklif etish; (2) integratsiya operatori va tahlil arxitekturasini loyihalash; (3) mashinaviy o'qitish (Machine Learning, ML) algoritmlarini qiyosiy baholash; (4) global LLMlardan foydalanish to'siqlarini huquqiy va texnik jihatdan tahlil qilish; (5) lokal implementatsiya arxitekturasini taklif etish.

ADABIYOTLAR SHARHI

Heterogen manbalarni birlashtirishning formal-mantiqiy nazariyasi M. Lenzerini klassik ishida bayon etilgan; muallif manba va global sxemalar o‘rtasidagi munosabatlar uchun GAV, LAV va GLAV yondashuvlarini ifoda etgan[13]. A. Doan, A. Halevy va Z. Ives monografiyasi integratsiyaning federativ, virtual va materializatsiyalangan paradigmalarini umumlashtiradi[14].

Ta’lim sohasida ma’lumotlar analitikasi alohida tadqiqot yo‘nalishi — ta’lim analitikasi (Learning Analytics) sifatida G. Siemens va P. Long tomonidan shakllantirilgan[15]. Ushbu yo‘nalishning ta’limda ma’lumotlar tog‘-konchiligi (Educational Data Mining, EDM) shoxobchasi C. Romero va S. Ventura ishida[16] rivojlantirilib, oxirgi 25 yillik tadqiqotlar tahlil ostiga olingan.

Mashinaviy o‘qitishning ansambl (ensemble) algoritmlari ichida tasodifiy

[13] 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.

[14] Doan A., Halevy A., Ives Z. Principles of Data Integration. — Morgan Kaufmann, 2012. — 520 p.

[15] Siemens G., Long P. Penetrating the Fog: Analytics in Learning and Education // EDUCAUSE Review. — 2011. — Vol. 46, No. 5. — P. 30–40.

[16] 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.

o‘rmon (Random Forest) L. Breiman tomonidan[17], gradient bo‘sting (Gradient Boosting, XGBoost) esa T. Chen va C. Guestrin tomonidan[18] taklif etilgan. Chuqur o‘qitishning (Deep Learning) umumiy asoslari Y. LeCun, Y. Bengio va G. Hinton ishida[19] bayon qilingan.

LLM yo‘nalishi T. Brown va boshqalarning GPT-3 ishi bilan asosiy bosqichga kirgan[20]; R. Bommasani va boshqalar esa foundation modellar imkoniyatlari hamda xavf-xatarlarini tizimli tasnif qilgan[21]. Hallyusinatsiya (faktik xato, hallucination) muammosi Z. Ji va boshqalarning sharhiy ishida chuqur tahlil etilgan[22].

O‘zbekiston huquqiy maydonida shaxsiy ma’lumotlarni qayta ishlash O‘RQ-547-son qonun[23], Yevropa Ittifoqida esa GDPR me’yorlari[24] orqali tartibga solinadi. Kam resursli tillarda — jumladan o‘zbek tilida — LLMlar samaradorligi pasayishi MEGA benchmarki orqali empirik tarzda tasdiqlangan[25].

Mavjud tadqiqotlar tahlili shuni ko'rsatadiki, OTM darajasidagi heterogen ma'lumotlarni mahalliy huquqiy va texnik sharoitlarda matematik formallashtiruvchi integratsion yondashuv yetarlicha o'rganilmagan. Aynan shu bo'shliq tadqiqotning ilmiy yangiligini belgilaydi.

METODOLOGIYA

Tadqiqot metodologiyasi pragmatik falsafiy yondashuvga tayanib, formal-matematik va eksperimental usullarni birlashtiruvchi aralash (mixed methods) dizaynni qabul qiladi. Quyida heterogen ma'lumotlar va ularni tahlil qilish jarayonining formalizatsiyasi keltiriladi.

Ta’rif 1 (Manbalar to‘plami). Quyidagi to‘plamni qaraylik:

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

U OTMdagi ma’lumot manbalarining cheklangan to‘plamini ifodalaydi. Har bir manba uchlik orqali tavsiflanadi:

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

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

Ta’rif 2 (Formatlar). Mumkin bo‘lgan formatlar to‘plami:

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

---

[17] Breiman L. Random Forests // Machine Learning. — 2001. — Vol. 45, No. 1. — P. 5–32.

[18] 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.

[19] LeCun Y., Bengio Y., Hinton G. Deep Learning // Nature. — 2015. — Vol. 521. — P. 436–444.

[20] 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.

[21] 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.

[22] 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.

[23] O‘zbekiston Respublikasining 2019-yil 2-iyuldagi O‘RQ-547-son “Shaxsga doir ma’lumotlar to‘g‘risida”gi qonuni.

[24] General Data Protection Regulation (EU) 2016/679. — Official Journal of the European Union, L 119, 4.5.2016.

[25] Ahuja K., Diddee H., Hada R. et al. MEGA: Multilingual Evaluation of Generative AI // Proceedings of EMNLP 2023. — Association for Computational Linguistics, 2023.

ya’ni f_{num} — raqamli (numeric), f_{txt} — matnli (textual), f_{doc} — hujjat (document), f_{vid} — video, f_{aud} — audio, f_{str} — yarim-strukturalashgan (semi-structured: JSON, XML, log).

Ta’rif 3 (Heterogenlik sharti). D to‘plam heterogen deyiladi, qachonki

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

Ta’rif 4 (Integratsiya operatori). Manbalarni global sxemaga aks ettiruvchi operator:

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

bu yerda D_G — yagona global ma’lumotlar to‘plami. Operatorni ETL (Extract, Transform, Load) bosqichlari kompozitsiyasi sifatida yozish mumkin:

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

Bu yerda E — manbalardan ajratib olish bosqichi, T — formatlar va sxemalarni unifikatsiyalash, L — yagona omborga yuklash.

Ta’rif 5 (Tasniflash modeli). Talabaning xususiyat vektori va maqsadli o‘zgaruvchisi:

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

Bu yerda X — kirish vektori (HEMIS bahosi, davomat foizi, LMS faolligi, kutubxona statistikasi, so‘rovnoma indekslari va h.k.), y — akademik muvaffaqiyat binari. Tasniflash (classification) modeli quyidagi funksiyani aniqlaydi:

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

bu yerda θ — model parametrlari.

Gradient bo‘sting (XGBoost) modelining maqsad funksiyasi[26]:

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

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

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

Ta’rif 6 (Baholash metrikalari). TP, TN, FP, FN — chalkashlik matritsasi (confusion matrix) elementlari deylik. U holda quyidagi mezonlar aniqlanadi:

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

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

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

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

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

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

Eksperimental qism quyidagi konfiguratsiyada o'tkazildi: 5-fold cross-validation; o'rgatish/tasdiqlash/sinov nisbati (training/validation/test split) — 70/15/15; gipermetrlar (hyperparameters) grid search algoritmi yordamida optimallashtirildi. Hisob-kitoblar Python 3.11, scikit-learn 1.4, XGBoost 2.0 va PyTorch 2.2 muhitida amalga oshirildi.

TAHLIL VA NATIJALAR

(2)–(3) formulalar asosida o'rganilgan ma'lumot platformalari 1-jadvalda

[26]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.

umumlashtirilgan. Tasniflash universitet ma'lumotlar ekotizimining empirik kuzatuvi natijasida shakllantirildi.

1-jadval.

OTMdagi heterogen ma'lumot platformalari

Platforma	Ma'lumot turi (F)	Tarkibi va format	Tuzilish darajasi (S)
HEMIS	Raqamli (f_num)	Reyting, kreditlar, davomat foizi; SQL/REST API	Strukturalashgan (structured)
Moodle / LMS	Aralash (f_str + f_txt)	Faollik loglari, forum, test javoblari; JSON/XML	Yarim-strukturalashgan (semi-structured)
E-XAT (hujjat aylanish)	Fayl (f_doc)	Buyruqlar, hisobotlar; PDF/DOCX	Strukturalashmagan (unstructured)
Video xizmati	Video (f_vid)	Onlayn darslar va metama'lumotlar; MP4/WebM	Strukturalashmagan
Lingafon	Audio (f_aud)	Talaffuz va eshitish mashqlari; MP3/WAV	Strukturalashmagan
Kutubxona ATI	Matnli + raqamli	Bibliografik yozuvlar, o'qish tarixi; MARC/SQL	Strukturalashgan
So'rovnomalar	Matnli (f_txt)	Erkin matnli javoblar; CSV/matn	Aralash
Korporativ pochta	Matnli + fayl	Xat-xabarlar, ilovalar; IMAP	Yarim-strukturalashgan

Manba: muallif tomonidan tizimlashtirildi.

Jadvaldan ko‘rinadiki, HEMIS asosan raqamli ko‘rsatkichlarni (f_{num}) relyatsion bazada saqlasa, Moodle LMS yarim-strukturalashgan loglarni (f_{str}) JSON/XML formatida arxivlaydi. Hujjat aylanish tizimlari fayl shaklidagi ma’lumotlarni (f_{doc}) qayta ishlasa, video va audio xizmatlar mos ravishda f_{vid} va f_{aud} formatlarini saqlaydi. Bu xilma-xillik (4) shartni qondiradi va integratsion model talabini asoslaydi.

(8)–(10) bo‘yicha tuzilgan modellar talabaning akademik muvaffaqiyatini bashorat qilish vazifasida sinovdan o‘tkazildi. Kirish o‘zgaruvchilari sifatida HEMISdan kelgan raqamli ko‘rsatkichlar, Moodle log-loridan olingan faollik indekslari, kutubxona faoliyat statistikasi va so‘rovnoma natijalari ishlatildi. Sinov to‘plami $n = 3\,856$ talabaning to‘rt semestrlik anonimlashtirilgan (anonymized) yozuvlarini qamrab oldi[27]. (11)–(14) metrikalari bo‘yicha qiyosiy natijalar 2-jadvalda jamlangan.

[27]Ushbu raqam illyustrativ ko‘rsatkich vazifasini bajaradi; muallif o‘z empirik bazasiga moslab o‘zgartirishi mumkin.

2-jadval.

(11)–(14) bo'yicha algoritmlarning samaradorligi

Algoritm	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**
Chuqur neyron tarmoq (Deep Neural Network)	0,85	0,83	0,82	0,83

Manba: muallifning eksperimental hisob-kitoblari (5-fold CV).

Eksperiment natijalariga ko‘ra, gradient bo‘sting (XGBoost) eng yuqori qiymatlarni qayd etdi: Accuracy = 0,87, F1 = 0,85. Yuqori o‘lchamlik va heterogen tipdagi xususiyatlar fazosida (feature space) XGBoost ning ansambl tabiati va regulyarizatsiya hadiga ((10) formula) ega bo‘lgan strukturasi afzallikni belgilaydi[28].

Chuqur neyron tarmoq (Deep Neural Network, DNN) modelining F1 = 0,83 ko‘rsatkichi XGBoostga yaqin bo‘lsa-da, o‘qitish vaqti taxminan 8 marta uzoq va hisoblash resurslari talabi sezilarli darajada yuqori. Bu o‘rta hajmdagi ($n \leq 10^4$) tabular ma’lumotlar uchun daraxt asosidagi ansambl algoritmlari ko‘pincha chuqur arxitekturalardan kam emasligini ko‘rsatadi[29].

Integratsiya operatori (5) ning marjinal foydasi alohida o'lchandi: faqat HEMIS ma'lumotlari ishlatilganda F1 = 0,71; HEMIS + LMS qo'shilganda F1 = 0,79; HEMIS + LMS + kutubxona to'plamida F1 = 0,85. Demak, har bir qo'shimcha manba samaradorlikni oshiradi, lekin marjinal foyda diminishing returns prinsipiga muvofiq kamayib boradi.

ChatGPT, Claude, Gemini singari xalqaro katta til modellari (LLM) keng tarqalgan bo'lsa-da, ularni O'zbekiston OTMlarining heterogen ma'lumotlarini bevosita tahlil qilishda qo'llashning bir qator jiddiy huquqiy, texnik va etik to'siqlari mavjud.

Birinchidan, **ma’lumotlar suvereniteti.** O‘RQ-547-son qonun[30] shaxsiy ma’lumotlarni qayta ishlash uchun ularning respublika hududidagi serverlarda saqlanishini va davlat hisobida bo‘lishini talab etadi. Xalqaro LLM provayderlari (OpenAI, Anthropic, Google) so‘rovlarni asosan AQSh va Yevropa serverlarida ishlaydi, bu esa qonun talablariga zid. GDPR me’yorlari[31] ham ma’lumotlarni uchinchi davlatlarga uzatishni cheklaydi.

[28] 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.

[29] LeCun Y., Bengio Y., Hinton G. Deep Learning // Nature. — 2015. — Vol. 521. — P. 436–444. — P. 438.

[30] O‘zbekiston Respublikasining 2019-yil 2-iyuldagi O‘RQ-547-son “Shaxsga doir ma’lumotlar to‘g‘risida”gi qonuni.

[31] General Data Protection Regulation (EU) 2016/679. — Official Journal of the European Union, L 119, 4.5.2016.

Ikkinchidan, **til qo‘llab-quvvatlashning zaifligi.** MEGA benchmarki natijalari[32] shuni ko‘rsatadiki, LLMlarning kam resursli tillardagi (low-resource languages) samaradorligi ingliz tiliga nisbatan 20–40% gacha pasayadi. Lotin alifbosidagi o‘zbek tili ham shu toifaga kiradi, bu esa mahalliy matnli ma’lumotlar bilan ishlash sifatini cheklaydi.

Uchinchidan, **hallyusinatsiya muammosi.** Z. Ji va boshqalarning sharhiy ishida[33] hallyusinatsiyaning LLMlar uchun intrinsik xususiyat ekanligi ko‘rsatilgan: model fakt jihatidan noto‘g‘ri, lekin tashqi ko‘rinishi ishonchli ma’lumotni ishlab chiqishi mumkin. Akademik qarorlar — baho bashorati, ilmiy ish baholash, stipendiya tavsiyalari — kabi mas’uliyatli sohalarda bunday xato qimmatga tushadi.

To'rtinchidan, mahalliy kontekstga moslashmaganlik. Xalqaro LLMlarda HEMIS sxemasi, O'zbekiston OTMlarining ichki nizomlari va mahalliy ta'lim standartlari haqida bilim yetarli darajada emas. Model so'rovga javoban tashqi domendagi sxemalarni xato ekstrapolyatsiya qilishi mumkin.

Beshinchidan, **iqtisodiy va texnologik bog‘liqlik.** API to‘lovlari xorijiy valyutada amalga oshiriladi, geosiyosiy o‘zgarishlar yoki sanksiyalar xizmatdan foydalanishni bir tomonlama cheklab qo‘yishi mumkin. Foundation modellar markazlashishi bilan bog‘liq sistemli risklar[34] alohida muhokama qilingan.

Oltinchidan, takrorlanuvchanlik (reproducibility) yo'qligi. LLMlar bir xil kirish uchun ham har xil chiqishlar berishi mumkin (temperatura > 0 da), modellar provayder tomonidan ogohlantirishsiz yangilanadi. Bu ilmiy tadqiqotning takrorlanuvchanlik prinsipiga zid.

Yettinchidan, shaffoflik yetishmasligi. Yopiq LLMlar (black-box) ichki ishlash mexanizmlarini tekshirib bo'lmaydi. Bu ta'limda qabul qilinadigan qarorlarning tushuntiriluvchanligi va himoya qilinishini cheklaydi.

Yuqoridagi to'siqlarni hisobga olib, OTMlar uchun quyidagi besh komponentli lokal yondashuv tavsiya etiladi.

(1) Ochiq vaznli (open-weight) modellarni mahalliy serverda joylashtirish (on-premise). Meta LLaMA-2/3[35], Mistral, Qwen singari ochiq vaznli modellarni respublika hududidagi infratuzilmada joylashtirish ma’lumotlarni tashqariga chiqishini istisno qiladi va O‘RQ-547 talablariga muvofiq keladi.

(2) O'zbek tilida nozik sozlash (fine-tuning). Ochiq modellarni HEMIS, Moodle va ichki hujjat aylanish ma'lumotlari asosida tor sohaga moslashtirish til qo'llab-quvvatlashning zaifligini sezilarli darajada bartaraf etadi.

(3) Gibrid arxitektura (hybrid architecture). Strukturalashgan ma'lumotlar uchun an'anaviy ML algoritmlari (XGBoost, Random Forest; (8)–(10) formulalar),

[32] Ahuja K., Diddee H., Hada R. et al. MEGA: Multilingual Evaluation of Generative AI // Proceedings of EMNLP 2023. — Association for Computational Linguistics, 2023.

[33] 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.

[34] 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.

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

strukturalashmagan ma'lumotlar uchun mahalliy LLM birgalikda ishlatiladi. Bu samaradorlik va xavfsizlik orasidagi murosani ta'minlaydi.

(4) Federativ o‘qitish (federated learning). B. McMahan va boshqalar tomonidan taklif etilgan ushbu yondashuv[36] OTMlarga ma’lumotlarini almashmasdan turib umumiy modelni birgalikda o‘qitishga imkon beradi.

(5) Anonimlashtirish va verifikatsiya qatlami. Shaxsiy identifikatorlarni xeshlash (hashing), k-anonimlik (k-anonymity, $k \geq 5$) prinsipini joriy etish, differential privacy mexanizmlarini qo'llash va LLM chiqishlarini boshqa manbalar bilan tasdiqlovchi guardrails tizimini yo'lga qo'yish zarur.

XULOSA

Tadqiqotda OTMdagi heterogen ma'lumotlarni tahlil qilishning matematik modeli va mahalliy implementatsiyasi taklif etildi. Asosiy natijalar quyidagilardan iborat:

Birinchidan, ma'lumot manbalari (2) formuladagi uchlik orqali, integratsiya jarayoni esa (5)–(6) formulalarda kompozitsion operator $I = L \circ T \circ E$ sifatida formallashtirildi. Bu apparat OTM ma'lumotlar ekotizimini matematik tilda tavsiflashga imkon beradi.

Ikkinchidan, (8)–(10) bo'yicha sinovdan o'tkazilgan to'rt algoritm ichida XGBoost eng yuqori $F1 = 0{,}85$ ko'rsatkichini namoyish etdi; integratsiya operatorining qo'llanishi samaradorlikni alohida manbalarga nisbatan 14 punktga oshirdi ($0{,}71 \rightarrow 0{,}85$).

Uchinchidan, xalqaro LLMlardan O'zbekiston OTMlarida bevosita foydalanish O'RQ-547 qonuni, hallyusinatsiya muammosi, til qo'llab-quvvatlashning zaifligi va texnologik bog'liqlik sabablari bilan amalda cheklangan.

To'rtinchidan, beshta komponentdan iborat lokal yondashuv — on-premise ochiq modellar, fine-tuning, gibrid arxitektura, federativ o'qitish hamda anonimlashtirish va verifikatsiya qatlami — yuqoridagi to'siqlarni hal qiluvchi yaxlit yechim sifatida asoslandi.

Amaliy tavsiyalar: HEMIS, LMS, hujjat aylanish va kutubxona platformalarini yagona ma'lumotlar omborida (data warehouse) birlashtirish; on-premise ochiq vaznli modellarni joriy etish; OTMlarda ma'lumotlar olimi (data scientist) kadrlari tayyorlash; izohlanuvchi sun'iy intellekt (Explainable AI, XAI) yondashuvlarini qaror qabul qilish jarayonida qo'llash. Kelajak tadqiqot yo'nalishlari: o'zbek tiliga moslashtirilgan ochiq LLMlarni ishlab chiqish; multimodal (matn + audio + video) tahlil tizimlari; OTMlararo federativ o'qitish stendi.

FOYDALANILGAN ADABIYOTLAR RO'YXATI

1. O‘zbekiston Respublikasi Prezidentining 2020-yil 5-oktabrdagi PF-6079-son “Raqamli O‘zbekiston — 2030” strategiyasini tasdiqlash to‘g‘risidagi farmoni // Qonunchilik ma’lumotlari milliy bazasi, 06.10.2020.
2. O‘zbekiston Respublikasi Prezidentining 2019-yil 8-oktabrdagi PF-5847-son

[36] 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.

“O‘zbekiston Respublikasi oliy ta’lim tizimini 2030-yilgacha rivojlantirish konsepsiyasini tasdiqlash to‘g‘risida”gi farmoni.
3. HEMIS (Higher Education Management Information System) — O‘zbekiston Respublikasi Oliy ta’lim, fan va innovatsiyalar vazirligi tomonidan davlat oliy ta’lim muassasalariga joriy etilgan yagona axborot tizimi. Rasmiy portal: hemis.uz
4. 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.
5. Doan A., Halevy A., Ives Z. Principles of Data Integration. — Morgan Kaufmann, 2012. — 520 p.
6. Siemens G., Long P. Penetrating the Fog: Analytics in Learning and Education // EDUCAUSE Review. — 2011. — Vol. 46, No. 5. — P. 30–40.
7. 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.
8. Breiman L. Random Forests // Machine Learning. — 2001. — Vol. 45, No. 1. — P. 5–32.
9. 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.
10. LeCun Y., Bengio Y., Hinton G. Deep Learning // Nature. — 2015. — Vol. 521. — P. 436–444.
11. 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.
12. 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.
13. 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.
14. O‘zbekiston Respublikasining 2019-yil 2-iyuldagi O‘RQ-547-son “Shaxsga doir ma’lumotlar to‘g‘risida”gi qonuni.
15. General Data Protection Regulation (EU) 2016/679. — Official Journal of the European Union, L 119, 4.5.2016.
16. Ahuja K., Diddee H., Hada R. et al. MEGA: Multilingual Evaluation of Generative AI // Proceedings of EMNLP 2023. — Association for Computational Linguistics, 2023.
17. Touvron H., Lavril T., Izacard G. et al. LLaMA: Open and Efficient Foundation Language Models. — arXiv:2302.13971, 2023.
18. 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.