

**SUN’IY INTELLEKT VA O‘ZBEK TILI: CHATGPT VA BOSHQA
NEYROTARMOQLARNING O‘ZBEK TILINI TUSHUNISH DARAJASI VA
MUAMMOLARI.**

Xaitmamatova Munisa

Denov tadbirkorlik va pedagogika instituti

Filologiya fakulteti Xorijiy til va adabiyoti: ingliz tili yo’nalishi talabasi

Annotatsiya: Ushbu maqolada sun’iy intellekt (SI) tizimlari, xususan, Katta Til Modellari (LLM – Large Language Models) bo’lmish ChatGPT (OpenAI), Gemini (Google) va Claude (Anthropic) kabi neyrotarmoqlarning o’zbek tilidagi matnlarni qayta ishlash, tushunish va generatsiya qilish qobiliyatlari tahlil qilinadi. Tadqiqot davomida o’zbek tilining agglyutinativ tabiati, morfologik murakkabligi va ikki xil yozuv (lotin va kirill) tizimining SI modellari samaradorligiga ta’siri o’rganildi. Natijalar shuni ko’rsatadiki, garchi LLMlar o’zbek tilida muloqot qilishda katta yutuqlarga erishgan bo’lsa-da, tokenizatsiya, uslubiy g’alilik, faktologik gallyutsinatsiya va madaniy kontekstni tushunishda jiddiy muammolar mavjud.

Kalit so’zlar: Sun’iy intellekt, o’zbek tili, ChatGPT, NLP (Tabiiy tilni qayta ishlash), LLM, agglyutinativ tillar, tokenizatsiya, mashina tarjimas.

KIRISH

XXI asrning uchinchi o’n yilligi sun’iy intellekt texnologiyalarining, xususan, generativ neyrotarmoqlarning keskin rivojlanishi bilan xarakterlanadi. 2017-yilda Google tadqiqotchilari tomonidan taqdim etilgan "Transformer" arxitekturasini va uning asosida yaratilgan Katta Til Modellari (LLM – Large Language Models) inson va mashina o’rtasidagi muloqot chegaralarini butunlay o’zgartirib yubordi. ChatGPT (OpenAI), Gemini (Google), Claude (Anthropic) va LLaMA (Meta) kabi tizimlar nafaqat axborotni qidirish, balki uni tahlil qilish, ijodiy matnlar yozish va dasturiy kodlar yaratish imkoniyatiga ega bo’ldi. Biroq, bu texnologik sakrash global miqyosda "raqamli tengsizlik"ni keltirib chiqarmoqda. Dunyo tillari sun’iy intellekt resurslari bilan ta’minlanganlik darajasiga ko’ra ikki guruhga bo’linib qoldi: "yuqori resursli" (high-resource) tillar (ingliz, xitoy, ispan) va "kam resursli" (low-resource) tillar. Afsuski, 50 milliondan ortiq so’zlashuvchiga ega bo’lgan o’zbek tili jahon axborot tarmog’ida hamon "kam resursli" tillar qatorida qolmoqda. ChatGPT, Google Gemini va Microsoft Copilot kabi tizimlar inson faoliyatining barcha jabhalariga kirib keldi. Bu jarayon global miqyosda ingliz, ispan yoki xitoy tillari kabi "yuqori resursli" (high-resource) tillar uchun ulkan imkoniyatlar ochgan bo’lsa-da, "kam resursli" (low-resource) yoki "o’rta

resursli" tillar hisoblangan o‘zbek tili uchun o‘ziga xos chaqiriqlarni yuzaga keltirdi. O‘zbek tili – turkiy tillar oilasiga mansub bo‘lib, o‘zining boy morfologiyasi, agglyutinatativ qurilishi (so‘z o‘zagiga qo‘shimchalar qo‘shish orqali yangi ma’no yasash) va gapda so‘z tartibining nisbiy erkinligi bilan ajralib turadi. Bu xususiyatlar asosan ingliz tili mantig‘i va ma’lumotlar bazasi asosida o‘qitilgan neyrotarmoqlar uchun jiddiy sinov hisoblanadi. Agar o‘zbek tili zamonaviy raqamli texnologiyalar va SI tizimlariga to‘laqonli integratsiya qilinmasa, kelajakda "raqamli o‘lim" yoki "texnologik izolyatsiya" xavfi ostida qolishi mumkin. Tadqiqotning maqsadi – mavjud mashhur LLMlarning o‘zbek tilini tushunish darajasini qiyosiy tahlil qilish, lingvistik xatoliklar tipologiyasini aniqlash va bu muammolarni bartaraf etish uchun ilmiy-amaliy tavsiyalar ishlab chiqishdir.

Ushbu tadqiqotning dolzarbligi shundaki, agar o‘zbek tili zamonaviy neyrotarmoqlarning "o‘quv bazasi"ga sifatli va to‘g‘ri integratsiya qilinmasa, quyidagi xatarlar yuzaga keladi:

1. Texnologik izolyatsiyadir. O‘zbek tilida so‘zlashuvchi aholi eng so‘nggi texnologik yechimlardan foydalanishda cheklovlar sezadi.

2. Tilning funksional torayishidir. Ilm-fan, biznes va IT sohalarida o‘zbek tili o‘rnini ingliz yoki rus tili to‘liq egallab olish xavfi ortadi.

3. Madaniy deformatsiyadir. Sun‘iy intellektning "inglizcha mantiq" asosida o‘zbekchani shakllantirishi milliy tafakkur va uslubga putur yetkazishi mumkin.

Shu bilan birga, o‘zbek tili va NLP (Tabiiy tilni qayta ishlash) muammolari O‘zbek tili turkiy tillar oilasiga mansub bo‘lib, o‘zining agglyutinatativ (qo‘shimchalar orqali so‘z yasash va turlash) xususiyati bilan ajralib turadi. Masalan, ingliz tilidagi "to my friends" birikmasi o‘zbek tilida bitta so‘z – "do‘stlarimga" shaklida ifodalanadi. Bu morfologik boylik neyrotarmoqlar uchun jiddiy qiyinchilik tug‘diradi. Mavjud LLMlar asosan ingliz tili ma’lumotlari asosida o‘qitilganligi sababli, ular o‘zbek tilidagi so‘z o‘zagi va qo‘shimchalar chegarasini aniqlashda (segmentatsiya) xatolarga yo‘l qo‘yadi. Bundan tashqari, O‘zbekistonda ikki xil yozuv tizimi (lotin va kirill) mavjudligi, internetdagi o‘zbekcha matnlarning (dataset) yetishmasligi va sifatining pastligi modellarning o‘rganish jarayonini sekinlashtirmoqda. Mazkur ishning asosiy maqsadi – yetakchi sun‘iy intellekt modellarining o‘zbek tili morfologiyasi, sintaksisi va semantikasini tushunish darajasini qiyosiy tahlil qilishdir.

Ushbu maqsadga erishish uchun quyidagi vazifalar belgilandi:

1. Zamonaviy LLMlarning (GPT-4o, Gemini 1.5 Pro, Claude 3.5 Sonnet) o‘zbek tilidagi matnlarni generatsiya qilish imkoniyatlarini baholash.

2. Tokenizatsiya jarayonida o‘zbek tili so‘zlarining noto‘g‘ri bo‘linish holatlarini aniqlash.

3. Modellarning o'zbek milliy madaniyatiga xos tushunchalarni (etnografizmlar, maqollar) anglash darajasini tekshirish.

4. Kelajakda o'zbek tili milliy korpusini yaratishda SI texnologiyalaridan foydalanish bo'yicha tavsiyalar ishlab chiqish.

2. METODLAR

Ushbu tadqiqotda empirik kuzatish, qiyosiy-tipologik tahlil va lingvistik eksperiment metodlaridan foydalanildi. Tadqiqot jarayoni to'rt bosqichda amalga oshirildi: ma'lumotlar to'plamini shakllantirish, modellarni tanlash, eksperiment o'tkazish va natijalarni baholash. Tahlil uchun bugungi kunda eng ilg'or va keng tarqalgan uchta Katta Til Modeli tanlab olindi:

- GPT-4o (OpenAI): Ko'p tilli (multilingual) imkoniyatlari eng yuqori baholangan model. Uning o'zbek tilidagi ma'lumotlar bazasi asosan Common Crawl (internetdagi ochiq matnlar) va Vikipediya ma'lumotlariga tayanadi.

- Google Gemini 1.5 Pro: Google qidiruv tizimi indeksatsiyasi va YouTube subtitrlari orqali katta hajmdagi o'zbekcha (jumladan, kirill va lotin yozuvidagi) ma'lumotlarga ega model.

- Claude 3.5 Sonnet (Anthropic): Matnning mantiqiy tuzilishi va xavfsizlik (safety) standartlariga yuqori e'tibor qaratadigan model.

Modellarning til bilish darajasini xolis baholash maqsadida jami 500 ta lingvistik birlikdan iborat maxsus sinov to'plami (benchmark dataset) shakllantirildi. To'plam quyidagi to'rt yo'nalishni qamrab oladi:

- Morfologiya (150 ta birlik): Ko'p bo'g'inli va murakkab qo'shimchali so'zlar (masalan: rivojlantirilayotganligidan, o'zlashtira olmaganlarimiz). Maqsad: Modelning so'z yasash va turlash qobiliyatini tekshirish.

- Sintaksis (150 ta gap): Sodda, qo'shma va murakkab gaplar. Maqsad: Gap bo'laklarining tartibi (Word Order) va egalik-kesim moslashuvini tekshirish.

- Semantika va Uslubiyat (100 ta matn parchasi): Rasmiy-idaho uslubi (ariza, qonun loyihasi) va badiiy uslub (parchalar). Maqsad: Modelning uslubiy farqlarni ajrata olishini aniqlash.

- Madaniy kod (100 ta savol): Maqollar, iboralar va O'zbekiston tarixi/madaniyatiga oid savollar. Maqsad: "Gallyutsinatsiya" (faktlarni to'qish) darajasini o'lchash.

Eksperiment davomida "Zero-shot" (namunasiz) va "Few-shot" (oz namunali) prompt texnologiyalaridan foydalanildi. Har bir modelga bir xil shartlar asosida quyidagi turdagi buyruqlar berildi:

- Tog'ridan-to'g'ri tarjima: "Quyidagi matnni o'zbek tiliga akademik uslubda tarjima qiling".

- Matn generatsiyasi: "Sun'iy intellektning ta'limdagi o'rni haqida 300 so'zlik insho yozing".

- Lingvistik tahlil: "Ushbu gapdagi grammatik xatolarni toping va tushuntiring".

Olingan natijalar ikki xil yondashuv asosida baholandi:

Avtomatlashtirilgan baholash: Tarjima sifati uchun BLEU (Bilingual Evaluation Understudy) va METEOR metrikalaridan qisman foydalanildi (matnlarning o'zbek tili morfologiyasiga mos kelish darajasini aniqlash uchun).

Ekspert baholash (Inson omili): 3 nafar o'zbek tili filologi va 2 nafar kompyuter lingvistikasi mutaxassisi natijalarni Likert shkalasi (1 dan 5 gacha) bo'yicha baholadilar:

5 ball: Mukammal (grammatik va uslubiy xatosiz).

4 ball: Yaxshi (kichik imloviy xatolar, ma'no to'g'ri).

3 ball: Qoniqarli (tushunsa bo'ladi, lekin uslub g'aliz).

2 ball: Yomon (ma'no buzilgan, jiddiy grammatik xatolar).

1 ball: Tushunarsiz (so'zlar yig'indisi yoki butunlay boshqa tilda).

Ushbu maqolada empirik kuzatish, qiyosiy tahlil va lingvistik eksperiment metodlaridan foydalanildi. Tadqiqot obyekti sifatida uchta asosiy Katta Til Modeli tanlab olindi:

1. GPT-4o (OpenAI): Hozirgi kunda eng ilg'or hisoblangan multimodal model.

2. Google Gemini (Pro 1.5): Google qidiruv tizimi ma'lumotlar bazasiga integratsiya qilingan model.

3. Claude 3.5 Sonnet (Anthropic): Katta hajmdagi matnlar bilan ishlashga ixtisoslashgan model.

Modellarning o'zbek tilini bilish darajasini aniqlash uchun ular 4 ta asosiy lingvistik mezon bo'yicha sinovdan o'tkazildi:

1. Morfologik va Sintaktik to'g'rilik: Murakkab qo'shimchalar zanjiriga ega so'zlarni (masalan, "kelayotganligidan", "tushuntirib berolmayapsizmi") tahlil qilish va gap tuzish qobiliyati.

2. Semantik (Ma'no) aniqlik: So'zlarning ko'p ma'noliligi va ko'chma ma'nolarini tushunish.

3. Uslubiy (Stilistik) moslik: Rasmiy, badiiy va so'zlashuv uslublarini farqlay olish.

4. Madaniy va Kontekstual tushunish: O'zbek maqollari, frazeologizmlari va milliy realiyalarni (masalan, mahalla, kelin salom, sumalak) to'g'ri talqin qilish.

Har bir modelga bir xil "prompt"lar (topshiriqlar) berildi va natijalar o'zbek tili filologiyasi qoidalari asosida baholandi.

Ushbu maqolada empirik kuzatish, qiyosiy tahlil va lingvistik eksperiment metodlaridan foydalanildi. Tadqiqot obyekti sifatida uchta asosiy Katta Til Modeli tanlab olindi:

GPT-4o (OpenAI): Hozirgi kunda eng ilg'or hisoblangan multimodal model.

Google Gemini (Pro 1.5): Google qidiruv tizimi ma'lumotlar bazasiga integratsiya qilingan model.

Claude 3.5 Sonnet (Anthropic): Katta hajmdagi matnlar bilan ishlashga ixtisoslashgan model.

Modellarning o'zbek tilini bilish darajasini aniqlash uchun ular 4 ta asosiy lingvistik mezon bo'yicha sinovdan o'tkazildi:

1. Morfologik va Sintaktik to'g'rilik: Murakkab qo'shimchalar zanjiriga ega so'zlarni (masalan, "kelayotganligidan", "tushuntirib berolmayapsizmi") tahlil qilish va gap tuzish qobiliyati.

2. Semantik (Ma'no) aniqlik: So'zlarning ko'p ma'noliligi va ko'chma ma'nolarini tushunish.

3. Uslubiy (Stilistik) moslik: Rasmiy, badiiy va so'zlashuv uslublarini farqlay olish.

4. Madaniy va Kontekstual tushunish: O'zbek maqollari, frazeologizmlari va milliy realiyalarni (masalan, mahalla, kelin salom, sumalak) to'g'ri talqin qilish.

Har bir modelga bir xil "prompt"lar (topshiriqlar) berildi va natijalar o'zbek tili filologiyasi qoidalari asosida baholandi.

XULOSA

Ushbu tadqiqot doirasida zamonaviy Katta Til Modellari (LLM) — GPT-4o, Google Gemini 1.5 Pro va Claude 3.5 Sonnet neyrotarmoqlarining o'zbek tilini tushunish, qayta ishlash va generatsiya qilish imkoniyatlari har tomonlama tahlil qilindi. O'tkazilgan lingvistik eksperimentlar va qiyosiy tahlillar natijasida quyidagi ilmiy-nazariy va amaliy xulosalar shakllantirildi: Sun'iy intellekt modellari o'zbek tilidagi morfologik va sintaktik to'siqlarni qisman yengib o'tgan bo'lsa-da, ular hali ham "mukammal so'zlashuvchi" darajasiga yetmagan. Xususan, GPT-4o modeli boshqalarga nisbatan yuqori natija ko'rsatgan bo'lsa-da, unda ham murakkab qo'shimchalar zanjirini (agglyutinatsiya) tahlil qilishda tokenizatsiya xatoliklari (so'zni noto'g'ri bo'laklash) saqlanib qolmoqda. Bu esa matnning ma'noviy butunligiga putur yetkazadi. Oliy ta'lim tizimida filologiya va jurnalistika yo'nalishi talabalariga sun'iy intellekt bilan ishlash ko'nikmalarini o'rgatish. Bu kelajakda neyrotarmoqlar yaratadigan matnlarni sifatli tahrir qilish uchun muhimdir.

Sun'iy intellekt — bu shunchaki vosita emas, balki tilning yashovchanligini belgilovchi yangi ekotizimdir. Agar biz o'zbek tilini ushbu ekotizimga to'g'ri integratsiya qila olmasak, tilimiz faqat maishiy muloqot darajasida qolib ketish va "raqamli o'lim"ga yuz tutish xavfi ostida qoladi.

Aksincha, bugungi kunda aniqlangan kamchiliklar (morfologik xatolar, ma'lumotlar yetishmasligi) bartaraf etilsa, sun'iy intellekt o'zbek tilining rivojlanishi, uning jahon axborot makonidagi o'rnini mustahkamlashi uchun misli ko'rilmagan imkoniyatlar eshigini ochadi.

FOYDALANILGAN ADABIYOTLAR

1. Rahimov, A. (2023). "Kompyuter lingvistikasi asoslari". Toshkent: Akademnashr.
2. Brown, T. et al. (2020). "Language Models are Few-Shot Learners". arXiv preprint.
3. O‘zbek tili izohli lug‘ati (2020). O‘zbekiston Milliy Ensiklopediyasi Davlat ilmiy nashriyoti.
4. OpenAI. (2024). GPT-4 Technical Report.
5. Mengliyev, B. (2022). "O‘zbek tili milliy korpusini yaratishning nazariy va amaliy masalalari". O‘zbek tili va adabiyoti jurnali.